WMG Service Systems Research Group
Working Paper Series

Measuring Individual Risk Attitudes when Preferences are Imprecise

Graham Loomes
Ganna Pogrebna

ISSN: 2049-4297
Issue Number: 09/13

About WMG Service Systems Group

The Service Systems research group at WMG works in collaboration with large organisations such as GlaxoSmithKline, Rolls-Royce, BAE Systems, IBM, Ministry of Defence as well as with SMEs researching into value constellations, new business models and value-creating service systems of people, product, service and technology.

The group conducts research that is capable of solving real problems in practice (ie. how and what do do), while also understanding theoretical abstractions from research (ie. why) so that the knowledge results in high-level publications necessary for its transfer across sector and industry. This approach ensures that the knowledge we create is relevant, impactful and grounded in research.

In particular, we pursue the knowledge of service systems for value co-creation that is replicable, scalable and transferable so that we can address some of the most difficult challenges faced by businesses, markets and society.

Research Streams

The WMG Service Systems research group conducts research that is capable of solving real problems in practice, and also to create theoretical abstractions from or research that is relevant and applicable across sector and industry, so that the impact of our research is substantial.

The group currently conducts research under six broad themes:

- Contextualisation
- Dematerialisation
- Service Design
- Value and Business Models
- Visualisation
- Viable Service Systems and Transformation

WMG Service Systems Research Group Working Paper Series
Issue number: 09/13
ISSN: 2049-4297
December 2013

Measuring Individual Risk Attitudes when Preferences are Imprecise

Graham Loomes
Warwick Business School, University of Warwick,
Gibbet Hill Road, Coventry, CV4 7AL, UK
E-mail: g.loomes@warwick.ac.uk

Ganna Pogrebna
Service Systems Group, Warwick Manufacturing Group,
University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK
E-mail: G.Pogrebna@warwick.ac.uk

Acknowledgement: Funding for this work came from a grant to Graham Loomes from the UK Economic and Social Research Council (grant no. RES-051-27-0248). Ganna Pogrebna acknowledges financial support from the Leverhulme Trust under the Early Career Fellowship scheme. An earlier version of this paper, titled "Can We Find Stable Measures of Individual Risk Attitudes?", was presented at the 2013 Annual Conference of the Royal Economic Society.
The authors are grateful to the attendees at that conference for many helpful suggestions. They also thank the participants at the Decision Research at Warwick (DR@W) Forum in the University of Warwick (UK), at the Department of Economics Research Seminar in the University of Sheffield (UK), at the 2011 Winter School in the University of Sydney (Australia) as well as at the 42nd Australian Conference of Economists in Perth (Australia) for many insightful comments. Finally, the authors very much appreciate the efforts of referees and an editor, which have greatly improved the final version of this paper.

If you wish to cite this paper, please use the following reference:

Loome G & Pogrebna G (2013) Measuring Individual Risk Attitudes when Preferences are Imprecise. *Ecomomic Journal,* forthcoming. Interim location: *WMG Service Systems Research Group Working Paper Series*, paper number 09/13, ISSN 2049-4297.

Measuring Individual Risk Attitudes when Preferences are Imprecise

By Graham Loomes and Ganna Pogrebna

Abstract

There is widespread interest in measuring risk attitudes and incorporating such measures into broader econometric analyses. We consider three elicitation procedures currently in use. We find considerable variability within – and even more, between – the results they produce. We suggest that this reflects the way that different instruments interact with imprecise underlying preferences. The short run implication is that such procedures need to be used with caution and are likely to be highly context-specific. The longer run implication is that adding 'white noise' to deterministic models is inadequate: we need to develop models that allow for imprecision and procedural variation.

Keywords: risk attitude, risk aversion, preference elicitation, procedural invariance

JEL classification: C81, C91, D81

1. Introduction

A number of recent studies into decision making under risk and over time have used various instruments to measure individual risk attitudes and discount rates. The idea is that such measures could be used to help explain and predict other decisions (such as choice of investment, insurance policy, pension scheme). This paper investigates the within- and between-procedural robustness of three such instruments when used to elicit risk attitudes. We find degrees of variability and disparity that are difficult to explain within the terms of any deterministic model, although it may be easier to understand at least some of the diversity and discrepancies if we allow that preferences are imprecise and that responses may, at least in part, be shaped by the procedures used to elicit them.

The three procedures which are the focus of this paper will be referred to as the *choice list* procedure, the *ranking* procedure and the *allocation* procedure. We shall describe each in detail in the course of the paper, but the essence of each is as follows.

The choice list (sometimes called multiple price list) method presents a table of binary choices designed so that as a respondent works through the table she can be expected to switch at some point from one 'side' to the other. When the choices are between risky alternatives, the switching point is assumed to be indicative of the individual's risk attitude.

The ranking procedure presents a set of options and asks the respondent to identify which option she ranks top. When applied to a set of risky prospects that have different combinations of spread and return, the idea is to identify the individual's risk attitude as reflected in her most-preferred balance between mean and variance.
The allocation procedure provides the respondent with a budget and allows her to distribute it between different state-contingent claims. When applied to risk, the chosen allocation, in conjunction with information about the rate of exchange between claims, should allow the individual's risk attitude to be inferred.

The data generated by any of the above procedures are normally interpreted with respect to expected utility theory (EUT) or some other deterministic theory, with the analysis allowing for some form of 'white noise' in people's responses. One needs to allow for within-person variability because, ever since the earliest experiments tried to elicit individuals' preferences, it has been known that if we present an individual with some set of decision tasks and repeat each question at several points within the same experimental session[1], we are likely to observe that individual giving a different answer to exactly the same question on at least some occasions (see Mosteller and Nogee, 1951, for an early example; a number of more recent studies are discussed in Bardsley et al., 2009, Chapter 7). So it has become standard practice to analyse responses on the basis that if an individual behaves according to Theory X, his/her true preferences are given by some deterministic 'core' parameters, with any particular observation liable to diverge from the true preference due to some

[1] Some experiments have involved repeating the tasks in different sessions separated by short periods of time (intended to be too short for individuals' preferences to have changed in any substantial way).

random noise, with the choice of error specification often being a matter of convention and convenience.

Such an approach raises two issues. One – which we note, but which is not the focus of this paper – is that different assumptions about the error term may produce quite different conclusions about the performance of different core theories – see Stott (2006) and Blavatskyy and Pogrebna (2010) for examples. The second issue, which is the motivation for this paper, is that the 'deterministic-core-plus-white-noise' formulation may be inappropriate and inadequate.

If the standard formulation *were* a reasonably good assumption, we should expect that responses to different questions within the same procedure should all be consistent with approximately the same core model and whatever risk attitude it entails, and that different procedures should yield much the same picture for any particular individual.

However, there is a substantial and wide-ranging body of evidence that casts doubt on such suppositions. In particular, there are many manifestations of 'framing' effects and failures of procedural invariance such that when ostensibly the same choice or decision task is presented in different formats or using different procedures, response patterns are *systematically* different. Kahneman and Tversky (2000) provide a collection of studies of such phenomena which are robust to replication.

One possible way of explaining both within-person variability and systematic between-procedure differences is to model responses to decision tasks not simply as the direct revelation of fully-formed 'true' preferences plus exogenous white noise but more as the result of some deliberative process whereby an individual draws on a substratum of possible values and preferences to construct a response.

Such an explanation is in the spirit of Simon's (1978) invocation to economists to develop models of procedural rationality by borrowing from neighbouring disciplines which have studied decision making processes. For example, there is a body of psychological and neuroscientific literature which models decisions as being arrived at after some neural 'accumulator' or 'sequential sampling' process (for a survey, see Otter et al., 2008; for an early and influential example applied to risky choice, see Busemeyer and Townsend, 1993). Instead of assuming, as deterministic models do, that an individual comes to a decision with a single precise set of subjective values, accumulator models suppose that it is as if past experiences have laid down distributions of such values. The process of reaching a decision is then modelled as if an individual samples repeatedly from these underlying distributions, building up subjective arguments for and against different options, with this mental sampling continuing until the balance of subjective feelings in favour of one option or another tips the individual to make a decision.

Models of this kind have several implications. First, they can accommodate the stochastic nature of decisions. It is easy to imagine cases where the relative advantages and disadvantages of competing options are such that the sampling process sometimes produces a balance in favour of one option and sometimes tips

the balance in favour of another, even when the options have been presented in exactly the same way. Such models can thereby provide an account of the variability of decisions when the same task is presented in exactly the same way on different occasions.

Second, such models can explain why an individual may take different amounts of time to process different decisions within the same format. For example, in binary choices where the alternatives are quite evenly balanced, it may require more sampling (i.e., take longer) to reach a decision and the probabilities of each being chosen may be in the vicinity of 0.5; whereas if one option is then improved relative to the other, it will be likely to be chosen more often and less time will be required to make that decision. Such a relationship between response times and choice probabilities is well-established (Jamieson and Petrusic, 1977; Moffatt, 2005), but it cannot easily be explained by a standard 'deterministic-core-plus-white-noise model'.

Third, such models may help to explain people's ability to recognise – up to a point, at least – their own uncertainty about their preferences. For example, Butler and Loomes (2007) asked individuals to respond to a series of binary choices where one option was held constant and the other was progressively changed. Every time the variable option changed, participants were asked to state which option they chose and whether they definitely preferred it or thought they preferred it but were not sure. From the perspective of standard deterministic models, it is hard to know what sense to make of such responses. Yet most participants found the question meaningful and were able to report 'imprecision intervals' where they were less than completely sure about their preferences. This is consistent with an accumulator process where the individual terminates sampling before he is completely confident about his decision.

Fourth, uncertainty about their preferences may make people susceptible to various procedural or contextual influences. Indeed, Butler and Loomes (2007) found that both ends of the reported imprecision intervals, as well as the point of switching between the two options, were liable to be influenced by various features of the procedure that standard economic models would regard as theoretically irrelevant. [2] However, it seems quite possible that the nature and framing of the decision task could systematically influence the sampling process and hence affect the patterns of response that result.

But how much do the possibilities outlined above really matter when it comes to eliciting measures of risk attitude? The answer depends on whether any interactions between the different procedures and people's deliberative processes result in systematic effects that are strong enough to undermine the generality and transferability of the measures we elicit. If the effects exist but are small and are counterbalanced by a number of other effects that offset or submerge them, we might operate on the basis that the combination of such effects can be adequately approximated by some standard error specification. On the other hand, if the effects

[2] For one subsample, the variable option started undesirable and progressively improved; for the other subsample, it started highly desirable and become progressively worse.

are powerful and highly context-dependent, we might want to be more cautious about supposing transferability from one context to another and we might put greater research effort into trying to develop procedural models which can accommodate such data.

To investigate these issues, we conducted an experiment designed to provide data about the amount of within-procedure variability and also about the extent of consistency or inconsistency between the three procedures under examination. To this end, we recruited 423 students from the University of Warwick who completed a set of 20 decision tasks which, between them, involved examples of all three procedures. Detailed information about the experiment and its implementation can be found online[3] and in the Supplementary Material, but the key features will be described more fully when each procedure is discussed in the relevant sections below.

Our results add considerably to the available evidence about the extent to which elicited measures of risk attitude, as conventionally defined, depend upon and vary with the particular questions asked within a given procedure. The results also suggest that even after allowing for within-procedure variability, there is substantial evidence that different procedures produce different patterns of response that are difficult – we think, impossible – to reconcile with a 'deterministic-core-plus-white-noise' approach.

The paper is structured as follows. In Sections 2, 3 and 4 we describe the experimental design relating to each elicitation procedure and discuss what the data may tell us about the extent to which we can rely on that particular procedure to deliver reasonably consistent measures of risk attitudes. [4] For all three procedures, the experimental payoffs are sums of money contingent on the realisation of single-stage random mechanisms where the likelihoods are clearly specified in a format comparable across procedures. Having looked at each procedure separately, Section 5 makes comparisons between them and discusses the nature and extent of the systematic differences we find. In Section 6, we consider implications which our results might have for future applications and directions of research.

2. The Choice List Procedure

2.1 Motivation

This procedure presents respondents with a table or list which constitutes an ordered series of pairs of options constructed in such a way that the point where an individual switches from one side of the table to the other is taken to identify a range within which the individual's point of indifference is located and from which a measure of the individual's risk attitude can be derived.

[3] Links to the experimental design are provided in Section 2.

[4] In order to focus on the main features of the experiment and the data generated, we provide the finer details of instructions and implementation in Appendix A, together with some additional analysis.

The choice list task has been used by many researchers for one purpose or another. An early example can be found in Cohen et al. (1987): they presented a particular lottery – e.g., a 0.25 chance of receiving 1,000 French Francs – and asked participants to consider a series of sure alternatives with increments of 50FF between them, in order to identify certainty equivalents. Tversky and Kahneman (1992, pp. 305-306) used a similar procedure in two stages, with the first stage offering seven widely spaced sure amounts against a given lottery and the second stage offering more finely gradated sums in the region where the first-stage switch had occurred. More recently, Holt and Laury (2002) – henceforth H&L – constructed tables where there were two-outcome lotteries on both sides of the table and where all payoffs were held constant while the probabilities were changed progressively in such a way that the preference could be expected to switch between one end of the table and the other. Such lists – or variants of them – have become popular in many studies since then.

In a world of reasonably robust deterministic preferences, this instrument could be expected to work quite well: if core preferences are consistent with EUT and if deviations in the form of switching a little too early or a little too late are due simply to white noise, two or three such choice lists should be sufficient to provide decent estimates of risk attitudes at the individual level and a good picture of the distribution of such measures in any sample.

However, there are two reasons for being cautious about operating on that basis: first, the doubt that EUT is the appropriate core model; and second, the imprecision of people's responses to such tasks and their vulnerability to procedural effects.

On the first issue, Cohen et al. (1987, p. 10) note that when choice lists are used to elicit certainty equivalents for a variety of lotteries, "the instability of risk attitudes is striking". Tversky and Kahneman (1992, Tables 3 and 4) provide similar evidence of how median responses vary from risk averse to risk seeking as the nature of the lotteries changes. Of course, such evidence does not by itself refute the 'deterministic-core-plus-white-noise' formulation: indeed, both sets of authors interpret their results as demonstrating the inadequacy of EUT as the core theory, proposing instead that some form of Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) provides a better core model, to which a conventional error term might well be added). An important lesson to be learned from these studies, however, is that taking just one or two choice list tables as a basis for eliciting measures of risk attitude and then extrapolating those measures to other choices is an unsafe way of proceeding. Notwithstanding such evidence, it has become common practice to rely on just one or two tables interpreted on the basis of EUT as the core theory.[5]

[5] For a recent example, see Brown and Kim (2013). In the manner of Tversky and Kahneman (1992) they use a two-step (coarser-finer) procedure but rely entirely on just one set of payoffs along the lines of Holt and Laury (2002) and assume EUT with constant relative risk aversion. Despite citing Cohen et al. (1987) as an early example of the choice list procedure, they make no reference to the concerns raised there about the use of EUT and the instability of EUT-based measures of risk attitude. No reference at all is made to Tversky and Kahneman (1992).

On the second issue, there is some evidence of imprecision in responses to choice lists. Cohen et al. (1987) provided the opportunity for respondents to identify values where they either considered the alternatives to be equivalent or else where they did not know which alternative they preferred (in which case they let the experimenter make the choice for them). If an individual identified at least two adjacent values in any one list, Cohen et al. registered a positive "indecision interval". With increments of 50FF, they found up to 10% of respondents reporting such indecision.

Dubourg et al. (1997) used lists to elicit willingness to pay for safety improvements, asking respondents to identify the largest amounts they felt sure they *would* be prepared to pay (call this WTP_{min}), the smallest amounts they were certain they would *not* pay (WTP_{max}) and, if those two amounts differed, the sum that they felt would make it hardest to decide whether to pay or not. They not only found substantial intervals between WTP_{max} and WTP_{min} but also noted that the size and position of those intervals was systematically affected by the ranges of values presented in the lists, sometimes to the extent that the average WTP_{min} elicited via a list with a larger range was higher than the average WTP_{max} elicited via a list with a smaller range.

More recently, Cubitt et al. (2013) elicited certainty equivalents and associated imprecision intervals using lists that were rather more fine-grained than those in Cohen et al. (1987) and they found imprecision to be pervasive, with an average of 87% of respondents identifying at least some interval. They did not test for the kinds of range effects reported by Dubourg et al. (1997) but they established the existence and persistence of imprecision across a broad range of lotteries of the kind widely used in incentivised experiments.

Of course, the fact that imprecision can be identified in choice list tasks does not necessarily mean that they are vulnerable to procedural effects. But Lévy-Garboua et al. (2012) consider several different variations of the H&L procedure and find that different ways of presenting the tables produce different patterns of risk attitudes and different degrees of internal inconsistency. So we set out to examine both the variability of estimated risk attitudes across different parameter sets and also the susceptibility of responses to a basic consistency test: namely, whether simply inverting the list makes any systematic difference to the distributions we infer.[6]

2.2 Design

We constructed five different lists, each constituting a separate decision task (henceforth, DT), varying the parameters of the choices in ways we shall explain shortly. We randomised our sample of participants so that about half saw the lists of choices ordered in one way – version 1 (henceforth V1) – while the other half (V2) saw exactly the same sets of choices, but presented 'upside down'.[7]

[6] By keeping the range of the table the same and also keeping all of the increments between rows the same and as uniformly spread as possible, we control for any 'range-frequency' effects of the kind identified in many experiments since Parducci (1965).

[7] Both versions of the experiment are available online via **https://columbia.qualtrics.com/SE/?SID=SV_9GH9H9WSptIYgJK (variation V1) and https://columbia.qualtrics.com/SE/?SID=SV_1Mte22jrFxXKYJu (variation V2).**

Figure 1: An Example of a Decision Task Display in the Choice List Procedure

Figure 1 shows how a list was presented. Here we show the final DT in the experiment, varying the lotteries on both sides in the style used by H&L. Our display was intended to make it easy for participants to see what the lotteries involved, how they compared with each other and how they changed from one row to the next. For the first two rows, the left-hand option offers the higher expected value (EV), while for the other eight rows the right-hand option offers the higher EV, with the difference between EVs increasing as we go down the list. Thus if someone chooses the right-hand option throughout, or else switches from left to right after the first row, she may be regarded as risk-seeking. If she switches between the second and third row, she may be judged approximately risk neutral. Switching below the third row is taken to signify risk aversion, with lower switching points indicating greater risk aversion.

Besides the DT shown in Figure 1, there were four other choice list tasks. Two of these were 'certainty equivalent' tasks where a particular lottery was held constant

on one side and on the other side the option of a sure sum of money was varied – in our cases, from £1 to £10 inclusive by £1 increments. For DT16, the fixed lottery offered a 0.6 chance of £10 and a 0.4 chance of 0; while for DT18, the fixed lottery gave a 0.2 chance of £18 and a 0.8 chance of £3.

The other two tasks took a 'lottery equivalent' form. In these cases, the sure sums of money on one side were held constant while on the other side a lottery offered two payoffs with the probabilities of the higher payoff varying by increments of 0.1 while the probability of the lower payoff reduced accordingly. For DT17, the sure amount was fixed at £8, with the alternative offering £18 with probability p and £3 with probability 1-p; while for DT19, all payoffs were £3 lower – that is, the fixed certainty of £5 was juxtaposed to (£15, p; 0, 1-p).

The incentive mechanism was straightforward: if a choice list question were selected to be the basis for payment, one row of the list would be picked at random and the participant would be given the option she had chosen from that pair and be paid according to how that option played out.

One feature of our experimental design worth bearing in mind when considering the results is that, within each procedure, participants were asked to answer all tasks of that kind and were then invited to review any earlier responses and make any adjustments they wished before confirming their complete set of answers to all DTs of the same kind. This was intended to give participants every opportunity to make their responses cohere in any respect they thought fit (but on the understanding that only one task from the experiment as a whole would form the basis of their payment for taking part).[8]

2.3 Results for the Choice List Procedure[9]

Table 1 reports the distributions of responses for each of the five DTs from the choice list procedure, categorising individuals by the number of times they chose the sure or safer option in each list[10] and distinguishing between V1 and V2. So the observations in the upper rows are those individuals exhibiting the most risk-seeking behaviour and lower rows represent progressively greater risk aversion. The average

[8] In many earlier studies that employed the kind of text format used by H&L, a substantial minority of participants exhibited inconsistency in the choice list procedure by switching more than once from one side of the table to the other. Crosetto and Filippin (2013) have collected a large meta dataset from 30 published studies with 4,726 participants. In this meta dataset, 16.3% of participants were inconsistent in the choice list procedure. Our display allowed us to significantly decrease the number of such inconsistencies: the proportion of inconsistent participants in our dataset ranged between 1.1% in DT19 to 5.6% in DT20.

[9] The results in this section and throughout the rest of the main text of this paper are based on 351 individuals out of a total of 423 who provided responses to the full set of tasks. In Appendix B we outline the criteria we used to exclude a participant's responses. Our criteria were quite demanding because we wanted to be able to reassure readers of this paper that none of the main conclusions could be driven by a number of outlier responses given by people who might not understand properly what they were being asked to do. Appendix C reports the analysis for the full sample of 423 individuals and shows that none of the conclusions are significantly altered. Furthermore, in Appendix D, we present an example of our analysis for a particular functional form EUT with CRRA and show that our results remain robust when we assume a particular functional form.

[10] The option that was the riskier lottery in nine rows might offer its high payoff with certainty in the tenth row, thereby dominating the sure/safer option in this case with a higher sure amount – but here this is counted as the choice of the riskier option.

number of sure/safer choices is shown at the bottom of each column. To provide a benchmark, the rows where a risk-neutral individual could be found are indicated by a darker colour.

Table 1 shows that the distributions of responses were indeed liable to be affected by something as seemingly arbitrary as which way up the choice lists were displayed. There is no particular effect for DT17 or DT19, but for DT16, DT18 and DT20, the differences are significant: in all three cases, Mann-Whitney tests reject the null hypothesis of no difference between the distributions at $p < 0.01$.

The contrasts between DT16 and DT18 are of particular interest, since each task involves a lottery with an EV of £6 being compared with sure sums ranging from £1 to £10 inclusive. Risk aversion might suggest that we should expect more safe choices when the lottery has higher variance, as in DT18, and this seems to be the case for V1 ($p < 0.05$); but turning the lists upside down reverses the difference, producing a significant tendency ($p < 0.01$) for V2 respondents to choose the sure option less often in DT18 than in DT16.[11] If responses can be shifted significantly by which way up a list is presented, we might want to be cautious about the extent to which precise numerical estimates of risk attitude parameters can be derived from just one or two list tasks and can then be used to 'control' for risk when analysing data generated in some rather different context.

Table 1: Distributions of Responses in DT16–DT20

	DT16 (£10, 0.6; 0, 0.4) vs £10 to £1		**DT17** (£18, p; £3, 1-p) vs £8		**DT18** (£18, 0.2; £3, 0.8) vs £10 to £1		**DT19** (£15, p; 0, 1-p) vs £5		**DT20** (£8, p; £5, 1-p) vs (£20, p; £1, 1-p)	
Sure/Safer	**V1**	**V2**	**V1**	**V2**	**V1**	**V2**	**V1**	**V2**	**V1**	**V2**
0	-	-	1	-	1	1	1	-	2	1
1	3	-	1	1	-	1	-	2	3	-
2	1	-	4	6	3	8	6	4	23	25
3	17	4	40	56	6	17	40	40	42	21
4	38	36	59	55	31	47	53	61	47	42
5	53	57	38	36	62	59	32	38	31	43
6	43	55	22	18	43	26	32	17	20	30
7	13	21	4	6	26	20	5	11	3	12
8	3	4	1	1	-	-	1	3	-	3
9	-	1	2	-	-	-	2	3	1	2
10	1	1	-	-	-	-	-	-	-	-
Average Sure/Safer choices	**4.98**	**5.41**	**4.31**	**4.13**	**5.22**	**4.73**	**4.40**	**4.44**	**3.88**	**4.48**

[11] The test conducted in these last two cases involves taking, for each individual, the difference between the number of times he/she chooses the sure option in each list and testing the null hypothesis that this difference is on average zero.

We deliberately use very general tests not requiring any particular functional form. Often, the responses to such choice lists are used in conjunction with some specific assumption about functional form – e.g., EUT with constant relative risk aversion (CRRA) – to generate mid-point estimates of risk attitude parameters. However, if such an exercise produces discrepancies, it is open to the objection that the wrong specification may have been chosen. To avoid this possible complication, we shall (in this and subsequent sections) ask instead how far each individual's response to one DT correlates with their response to a different DT *relative to* other members of the same sample who were shown the list displayed in the same format.

If 'attitude to risk' is an individual-level characteristic, we might expect (at least within a particular kind of elicitation procedure) that those individuals who exhibit more (less) risk aversion than others in one DT would also be likely to exhibit more (less) risk aversion than those same other people in a different DT of the same kind. We examine this hypothesis in the context of the five choice lists by ranking individuals according to their switching points within each list, supposing that if individual X is more risk averse than individual Y, X will tend to switch at lower points in Table 1 than Y.

Tables 2 and 3 show the rank correlation coefficients for each pair of tasks within each version of the experiment. In most comparisons, the correlations are positive and statistically significant at the 1% level, which suggests that there is some broad case for thinking of some people as more or less risk averse than others. However, the coefficients vary a good deal: they are highest for the two lists we might regard as most similar – DT17 and DT19, where DT17 is the same as DT19 except with £3 added to each payoff – but are often quite low for other pairs where there are more differences: for example, between DT16 (where a lottery with a possibility of a zero payoff stays fixed while the alternative sure amounts vary) and DT17 (where the sure amount is fixed while the alternative lotteries vary the probabilities of two strictly positive payoffs).

Table 2 Spearman Rank Correlation Coefficients, Choice List Procedure, V1

	DT17	**DT18**	**DT19**	**DT20**
DT16	0.110	0.202**	0.316***	0.344***
DT17		0.234**	0.645***	0.370***
DT18			0.249**	0.306***
DT19				0.560***

** significant at 0.01; *** significant at 0.001 level

Table 3: Spearman Rank Correlation Coefficients, Choice List Procedure, V2

	DT17	DT18	DT19	DT20
DT16	0.284***	0.289***	0.313***	0.447***
DT17		0.297***	0.559***	0.464***
DT18			0.225**	0.204**
DT19				0.493***

** significant at 0.01; *** significant at 0.001 level

If a method of obtaining measures of risk attitude is to be given much credence, one might expect it to do better in ranking individuals consistently within the narrow domain of lotteries involving no more than two payoffs.

This conclusion might seem rather more pessimistic than the one reached by Andersen et al. (2008). In their study, they asked each participant to respond to four choice list tasks and declared themselves (p. 591) content to use these responses as the basis for estimating CRRA coefficients. However, all four of their lists shared the H&L format with payoffs not varying greatly across the lists, so it would not be too surprising if they found enough consistency over that range to satisfy themselves. Andersen et al. (2008) did not explore how well their estimates extended to other structures within the choice list format, as we did; nor did they examine their transferability to other kinds of tasks. This is an issue to which we now turn.

3. The Ranking Procedure

3.1 Design and Motivation

Our ranking task is a variant of a procedure used in Binswanger (1980, 1981) and further developed by Eckel and Grossman (2002) who used a multiple choice task where each participant was asked to select the most preferred lottery out of a set of five 50-50 lotteries with different payoffs and expected values. These lotteries were shown as rows in a table, with the top row offering a sure amount and subsequent lotteries keeping the probabilities fixed while progressively increasing the EV and the spread of the payoffs. Those respondents who were most risk averse could opt for the certainty, but less risk averse individuals were expected to prefer lotteries lower down the table, with the least risk averse (and all risk neutral and risk seeking) individuals opting for the lottery in the bottom row, which offered the highest EV.

We modified the Eckel and Grossman (2002) – henceforth E&G – method in two ways. First, we increased the number of lotteries in the table from five to six so that there was no longer a 'middle' item and there was room for finer differentiation of attitudes. Second, instead of asking participants just to select one lottery out of six, we asked them to rank all lotteries from the most preferred (rank 1) to the least preferred (rank 6). This allowed us not only to do what E&G did (by looking at the

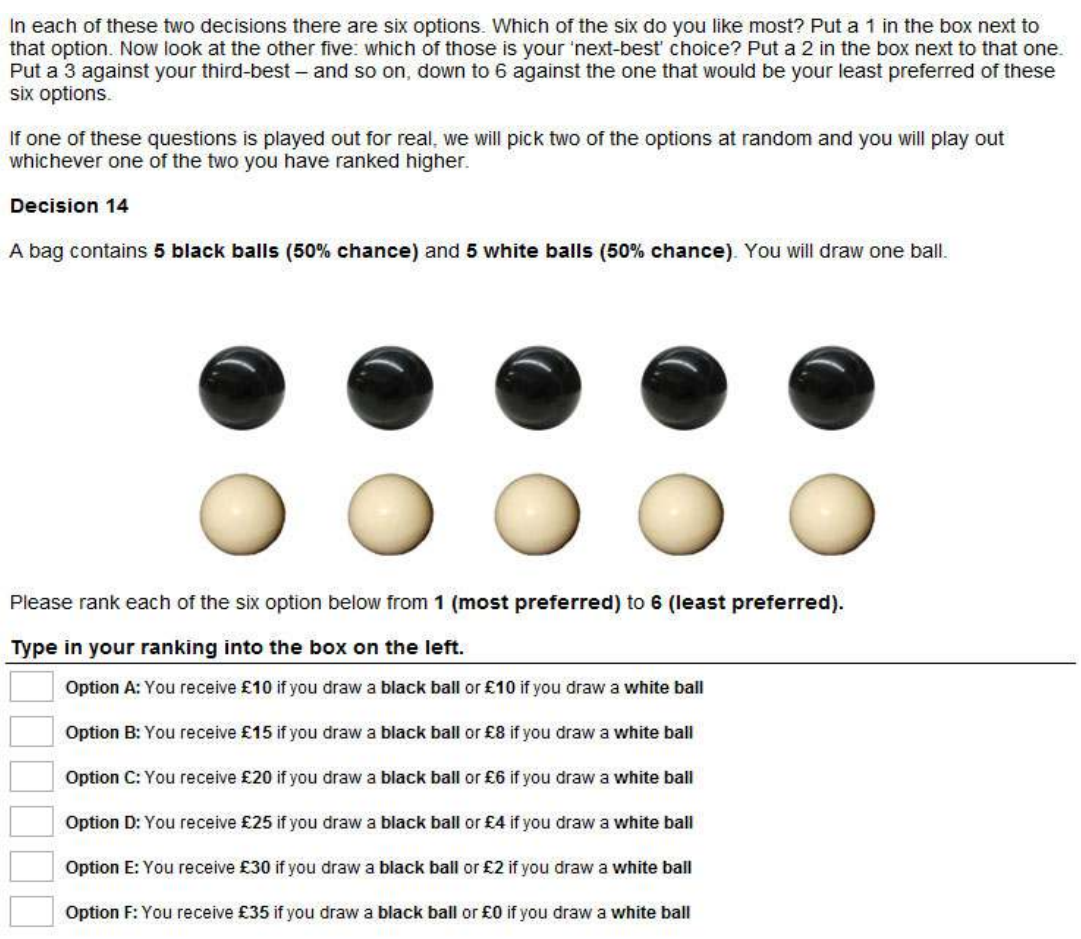
top choices), but also to see how far individuals' responses look consistent with the 'single-peakedness' property often assumed in the modelling of preferences[12]. By asking each participant to undertake the task for two different sets of six lotteries, we aimed to explore how sensitive responses were to changes in the parameters.

In order to encourage respondents to think carefully about the whole ranking, they were told that if one of the questions were selected to be the basis for payment, two of the six options would be picked at random and the participant would be given whichever of those two options she had ranked higher and be paid according to how that option played out.

In a deterministic world, ranking tasks are extensions of binary preferences (so long as transitivity holds) and the order in which the options are processed should make no difference to the final result. However, if preferences are imprecise, the order in which the options are processed or the range of the options or the intervals between each alternative may affect the final ranking.

The first of the ranking DTs was the same for every participant. It is shown in Figure 2 below.

Figure 2: Decision Task Display in the First Ranking Task



[12] This latter issue is discussed more fully in Appendix B: in the main text, we shall focus just on the top-ranked options as in Binswanger (1980, 1981) as well as in Eckel and Grossman (2002).

However, the second ranking task differed between 'treatment' variations, as shown in Table 4. For the V1 subsample, the pairs of payoffs in each row were the same as in the first ranking task, but the Black:White probability ratio became 0.3:0.7 rather than 0.5:0.5, so that the EV rose much more slowly from top row to bottom row (increasing from £10.00 to £10.50 in increments of £0.10 rather than from £10.00 to £17.50 in increments of £1.50). The conventional wisdom here is that only those who are risk seeking or risk neutral (or at most just very slightly risk averse) will rank F higher than A, and the only people who should place B, C, D or E first are those with sufficient risk aversion to turn down F but not enough risk aversion to prefer A.

Table 4: Parameters of the Various Ranking Task Lotteries

Treatment variation	Decision task	Lottery	Outcome Black	Outcome White	EV
			Prob = 0.5	Prob = 0.5	
V1, V2	DT14	A	10	10	10.0
		B	15	8	11.5
		C	20	6	13.0
		D	25	4	14.5
		E	30	2	16.0
		F	35	0	17.5
			Prob = 0.3	Prob = 0.7	
V1	DT15	A	10	10	10.0
		B	15	8	10.1
		C	20	6	10.2
		D	25	4	10.3
		E	30	2	10.4
		F	35	0	10.5
			Prob = 0.7	Prob = 0.3	
V2	DT15	A	10	10	10.0
		B	12	8	10.8
		C	14	6	11.6
		D	16	4	12.4
		E	18	2	13.2
		F	20	0	14.0

For the V2 subsample, the Black:White probability ratio was 0.7:0.3 and the payoffs in each row were different divisions of £20, from £10:£10 in the top row to £20:0 in the bottom row, with EVs increasing from £10.00 to £14.00 in increments of £0.80. As we shall see later, this latter table was intended to provide a direct comparison with one of the allocation tasks to be discussed in Section 4. But it is also the case that the relationship between the EVs and the variances of the bets is broadly comparable with that for DT14, so that, to the extent that measures of risk attitude reflect this relationship, we should expect to see those choosing A and B in DT14 being spread over A, B and C in DT15, with those choosing C and D in DT14 opting for

D, E and F in DT15 and those picking E and F in DT14 being expected to opt for F in DT15.

3.2 Results for the Ranking Procedure

As with the choice list procedure, participants could scroll between both ranking procedure tasks and adjust their responses to either one in the light of their answers to the other. Again, this was intended to allow the opportunity for whatever kind of consistency the individual wished to achieve.

Table 5 reports the numbers of individuals in the V1 subsample putting each option at the top of their ranking.

Table 5: 1st Choices in DT14 and DT15, V1 subsample

DT14 1st Choice		DT15 1st Choice	
A: £10 for sure	70	118	A: £10 for sure
B: £15, 0.5; £8, 0.5	28	22	B: £15, 0.3; £8, 0.7
C: £20, 0.5; £6, 0.5	23	9	C: £20, 0.3; £6, 0.7
D: £25, 0.5; £4, 0.5	13	2	D: £25, 0.3; £4, 0.7
E: £30, 0.5; £2, 0.5	2	2	E: £30, 0.3; £2, 0.7
F: £35, 0.5; £0, 0.5	36	19	F: £35, 0.3; £0, 0.7

So the direction of movement is much as we should have expected, although of the 70 who chose the sure option in DT14, 12 moved away to a riskier position in DT15, despite those options offering much poorer gains in return for the extra risk. The Spearman rank correlation coefficient of 0.331 is significant at the 1% level, but is arguably quite low (although the preponderance of A choices in both tasks – but especially in DT15 – might limit the sensitivity of the coefficient).

For the V2 subsample, the first choice data are shown in Table 6 below. Here too, the overall distribution of first choices moves in the expected direction. The Spearman rank correlation between DT14 and DT15 is 0.495, which is significant at the 1% level and higher than most of the correlations reported so far. However, when we look at individual behaviour, there were indications that both A and F may have been somewhat oversubscribed in DT14. The 77 who opted for A in DT14 (and therefore turned down the mean-variance trade-off in B or beyond) should not on that basis have opted for anything riskier than B in DT15; yet 39 of them opted for C, and another 17 opted for D, E or F in roughly equal numbers. At the other end of the table, all 45 with sufficiently little aversion to risk that they top-ranked F in DT14 might also have been expected to select F in DT15. However, only 30 did so, with another 13 opting for A, B or C. This would be consistent with the possibility that imprecise preferences, in conjunction with procedures which start at one end or other of the table, may result in disproportionate numbers in those end categories.

Table 6: 1st Choices in DT14 and DT15, V2 subsample

DT14 1st Choice		DT15 1st Choice	
A: £10 for sure	77	19	A: £10 for sure
B: £15, 0.5; £8, 0.5	22	15	B: £12, 0.7; £8, 0.3
C: £20, 0.5; £6, 0.5	26	67	C: £14, 0.7; £6, 0.3
D: £25, 0.5; £4, 0.5	5	22	D: £16, 0.7; £4, 0.3
E: £30, 0.5; £2, 0.5	4	16	E: £18, 0.7; £2, 0.3
F: £35, 0.5; £0, 0.5	45	40	F: £20, 0.7; £0, 0.3

One noteworthy feature of the DT15 distribution in Table 6 is the mode at option C. It is possible that the popularity of that option in this task is due in part to factors that we shall see at play in the allocation procedure, which provided the first thirteen tasks in the experiment. We focus on that procedure next.

4. The Allocation Procedure

4.1 Design and Motivation

Our experimental examination of the allocation procedure was built on a design used in Loomes (1991). Our respondents were given a fixed total sum of money (£20) and were invited to allocate it in any way they wished between different possible states of the world that are contingent upon the outcome of a well-defined random mechanism. In this case the random mechanism was a bag of 10 coloured balls.

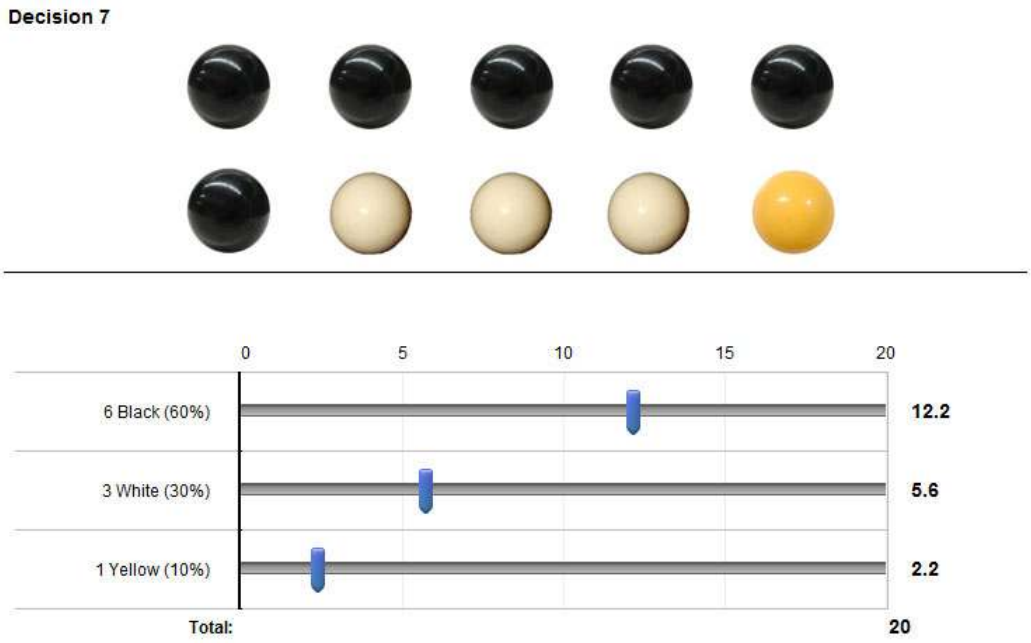
There were thirteen different distributions of up to three colours, as shown in Table 7 below, where each distribution constituted a different DT.

Table 7: Design of the Allocation Task

Decision Task	Probability of the outcome =		
	"Black ball"	"White ball"	"Yellow ball"
DT1	0.9	0.1	0.0
DT2	0.8	0.2	0.0
DT3	0.8	0.1	0.1
DT4	0.7	0.3	0.0
DT5	0.7	0.2	0.1
DT6	0.6	0.4	0.0
DT7	0.6	0.3	0.1
DT8	0.6	0.2	0.2
DT9	0.5	0.5	0.0
DT10	0.5	0.4	0.1
DT11	0.5	0.3	0.2
DT12	0.4	0.4	0.2
DT13	0.4	0.3	0.3

Each task was implemented via a computer display such as the one shown in Figure 3 below. In each case, participants indicated how much of the £20 they wished to allocate to each colour by moving the sliders, thereby setting the amounts at the right hand end of each slider to the nearest £0.10. It was explained that if one of these questions were selected to be the basis for the individual's payment, the relevant mix of coloured balls would be put into an opaque bag. The respondent would then draw one ball at random from that bag and receive whatever sum of money she had allocated to that colour.

Figure 3: Example of Display for an Allocation Procedure Decision Task



Participants moved from one DT to the next by scrolling down, and after making their thirteenth allocation they were invited to scroll back up to earlier answers and adjust them if they wished, in order to give them every opportunity to reflect upon their decision strategies and give the responses that they were content to have as the basis for payment. Once participants were satisfied with all their responses to the allocation DTs, they confirmed them as a whole.

How might a standard deterministic model of state-contingent claims expect participants to behave? If respondents' subjective values for the payoffs are functions only of the amounts on offer in the experiment, they will divide the £20, allocating X to Black, Y to White and Z to Yellow so that the marginal utilities of X, Y and Z are inversely proportional to the relative probabilities of success – i.e., in the example shown in Figure 3, so as to achieve u'(Z) = 3u'(Y) = 6u'(X).[13] So those who are

[13] A standard formulation of EUT, where utility is a function of wealth, would suppose that an individual who comes to the experiment with existing wealth W would try to allocate the £20 between X, Y and Z so as to make the ratios of the marginal utilities of W+X, W+Y and W+Z equal to the inverse of the ratios of the probabilities of the different colours being drawn. However, much experimental evidence suggests that many respondents act as if they are not integrating payoffs with W

risk seeking, risk neutral or only slightly risk averse will set X = £20 and Y = Z = 0, producing an allocation with the highest available expected value (here, £12). Those who are extremely risk averse will set X = Y = £6.70 and Z = £6.60, giving an expected value of £6.69. Intermediate degrees of risk aversion result in allocations somewhere between the two, with lower EVs correlating with higher degrees of risk aversion. If we were to assume a particular functional form of EU such as CRRA, the 13 allocation decisions would allow us to estimate the risk attitude parameter which best organises a particular individual's responses.[14]

Recently, researchers have used different variants of the general idea of asking respondents to make allocation decisions. For example, Choi et al. (2007) asked their participants to distribute their budget across two states of nature whose probabilities of occurrence were fixed (either at ⅓ : ⅔ or at ½ : ½) but where the relative costs of state-contingent claims were varied. They concluded that many participants in the population could be divided into three main groups: (a) extremely risk averse types who always opt for a safe portfolio choice; (b) participants whose behaviour is consistent with risk neutrality and (c) participants who do not fall within (a) or (b) but whose decisions were approximately compatible with a simple proportionality heuristic. Not all participants fell cleanly into one or other category: some seemed to jump from one to another; others seemed to approximate one of the categories but with modifications.

More recently, Andreoni and Sprenger (2012a) asked respondents to allocate tokens between different payoff dates, with earlier-paying tokens usually being worth less than later-dated tokens[15] and with the sizes of those differences and the lengths of time between payoff dates being varied. A proportionality heuristic was not so easily available in this task and about 70% of allocations were corner solutions, either allocating all 100 tokens to the earlier date or else having all 100 tokens pay out at the later date, with 24 of the 97 respondents allocating all 100 tokens to the later date on every single occasion when later-dated tokens paid more, however small the difference and whatever the time delay involved.

In a companion paper, Andreoni and Sprenger (2012b) compared cases where the time-dated payments were certain and cases where there was some probability less than 1 that they would be received: here certain payments seemed to be treated quite differently from risky payments, and although corner solutions still constituted more than a quarter of all observations in the risky scenarios, this was very different from the 80% of corner solutions in the certain scenarios in this experiment.

These results are incompatible with standard deterministic-core-plus-white-noise models, but there is little discussion about the extent to which they might be artefacts of interactions between imprecise preferences and the allocation procedure. However, Cheung (2013) compared their allocation procedure with a choice list procedure: regarding the dichotomy between certain and uncertain

---

but are focusing mainly or solely on gains or losses relative to their *status quo* level of wealth – see Rabin (2000).

[14] See Appendix D for the calculations of CRRA coefficients.

[15] Except for one of the 45 variants, where token values were the same while later-dated tokens involved a delay of 70 days.

payments, he states in his Abstract that "the effect disappears completely when a multiple price list instrument is used . . .".

Despite their possible susceptibility to rules of thumb as a way of coping with imprecise preferences, both Choi et al. and Andreoni and Sprenger make strong claims for the usefulness and greater plausibility of the estimates derived from their allocation tasks. In a recent paper, Charness and Gneezy (2012) have also advocated the use of an allocation procedure for measuring risk attitudes. Additional support for this kind of procedure comes from Hey and Pace (2011, p. 2) who assert that such tasks "are more informative than pairwise choice questions and probably more reliable than reservation price questions and thus more able to detect true preferences". However, with the exception of Cheung (2013), we are not aware of any systematic comparison that has been conducted between this procedure and other methods of eliciting risk attitudes, so an investigation of the properties of such a task should be of interest.

4.2 Results for the Allocation Procedure[16]

To give some initial picture of the way people responded to the allocation task, Table 8 shows the distributions of amounts allocated to Black for DT1, DT2, DT4 and DT6, the four simplest cases where there were only Black and White balls and where Prob(Black) > 0.5. The amounts allocated to Black are shown by the row labels and the columns show the numbers of individuals falling into each row for the four different DTs. Since respondents had strong inclinations to give answers rounded to the nearest whole pound, every other row shows whole-pound responses, while the rows in between capture responses involving other multiples of £0.10.[17]

Going to the bottom row, we see that in DT1, 156 individuals allocated all £20 to Black. Reading along that bottom row shows that the numbers allocating all £20 to Black falls markedly as the probability of Black falls, leaving just 16 in this category for DT6. Of these, 6 displayed at least some degree of risk aversion in other DTs, leaving just 10 who allocated all £20 to the highest probability events in every DT. Thus at most there are 10 out of 351 participants who display behaviour consistent with risk neutrality or risk seeking as judged on the basis of this procedure.

[16] For the allocation procedure, there was no difference between V1 and V2 – all participants saw exactly the same decision tasks in the same order until they reached DT15, as described in the previous Section.

[17] From £10 to £20 inclusive, there are 101 multiples of £0.10, of which just 11 are whole pounds; yet in Table 8, whole pounds account for 89% of responses. The slider mechanism did not make it easier to give whole pound responses than any other (apart, perhaps from the £20 endpoint), so this pattern strongly suggests a high degree of deliberate rounding in most people's responses – arguably a further reflection of the imprecision of people's preferences.

Table 8: Allocation to Black in DT1, DT2, DT4 and DT6

£ in Black	DT1 0.9 : 0.1	DT2 0.8 : 0.2	DT4 0.7 : 0.3	DT6 0.6 : 0.4
10	6	6	14	71
	-	1	4	11
11	-	-	-	30
	1	-	3	6
12	1	-	29	154
	-	1	8	20
13	1	5	21	7
	-	2	7	5
14	-	3	111	6
	-	1	10	-
15	26	57	71	16
	2	13	13	4
16	6	104	10	4
	1	7	2	-
17	18	18	10	1
	10	8	1	-
18	96	31	3	-
	7	1	1	-
19	16	8	-	-
	4	3	-	-
20	156	82	33	16

For DT2, DT4 and DT6, the modal response involves dividing the £20 in exact proportion to the probabilities, while for DT1 this is the second most popular response. In fact, 55 (15.7%) of the 351 respondents divided the £20 in exact proportion to the probabilities in *every one* of the 13 allocation tasks; and 37.4% of all responses exhibited such proportionality[18]. One interpretation is that a substantial minority of respondents' utility functions are logarithmic (or approximately so) in gains. A rather different interpretation is that many respondents who have somewhat imprecise preferences are content to adopt a simple proportionality heuristic which has broadly appealing properties – it involves putting more money on higher-probability events while not staking everything on just one – and which just happens to operate in the same way as logarithmic utility. If the first interpretation is correct, we should expect these individuals to exhibit behaviour consistent with a logarithmic utility function in the ranking and choice list tasks; if the second interpretation is closer to the mark, logarithmic utility will fail to predict behaviour in those other tasks where the same heuristic is not so readily available. We shall return to this issue in Section 5.

Those individuals exhibiting proportionality are a substantial minority, but a minority nevertheless. In order to form some view about the extent to which the allocation task provides a measure of relative risk attitude across the sample as a whole while

[18] This last figure excludes DT9, where there were 5 Black and 5 White balls and where 322 divided the £20 equally: including these takes the overall average to 41.6%.

avoiding any specific functional form, we rank individuals within each DT according to the EVs of their chosen allocation. That is, we suppose that if individual G is more risk averse than individual H, G will prefer an allocation in each task which entails a smaller spread of payoffs and a lower EV than the allocation selected by H.

Table 9 shows the Spearman rank correlation between every pair of DTs (except DT9 where all EVs were necessarily the same). Although all of these correlation coefficients are statistically significant at the 1% level, there is considerable variability, from 0.843 down to 0.187. As we found with the choice list tasks, the correlations tend to be highest between those tasks that are adjacent and/or with the most similar parameters, and tend to fall for pairs that are further apart and/or more different. If we can only make rather limited inferences from one allocation task to another, how much more cautious should we be about inferences to other kinds of decision task? That is the question addressed in the next Section.

Table 9: Spearman Rank Correlation Coefficients, Allocation Procedure

	DT2	DT3	DT4	DT5	DT6	DT7	DT8	DT10	DT11	DT12	DT13
DT1	0.744	0.733	0.513	0.621	0.187	0.513	0.324	0.640	0.397	0.414	0.232
DT2		0.843	0.672	0.713	0.365	0.618	0.479	0.582	0.459	0.500	0.300
DT3			0.634	0.725	0.354	0.638	0.510	0.599	0.454	0.434	0.286
DT4				0.817	0.463	0.588	0.512	0.480	0.436	0.427	0.326
DT5					0.457	0.728	0.573	0.619	0.501	0.551	0.359
DT6						0.588	0.662	0.266	0.353	0.321	0.258
DT7							0.730	0.645	0.571	0.533	0.360
DT8								0.391	0.537	0.491	0.367
DT10									0.553	0.532	0.278
DT11										0.606	0.371
DT12											0.417

5. Comparisons Between Procedures

5.1 Previous Comparisons Within and Between Procedures

Several studies have compared elicited risk attitudes across procedures, with mixed results. Dave et al. (2010) compare H&L and E&G procedures and correlate obtained results with the measure of numerical skills. They conclude that EUT with CRRA fits the data from both tasks equally well in the subsample of participants with relatively low numerical skills, but for those with high numerical skills, EUT with CRRA fits the data better in the H&L task than in the E&G task. Harrison and Rutstrom (2008, p.82) compare both methods with a third based on separate binary choices and derive CRRA estimates from all three. They note that the H&L procedure exhibits a statistically significant order effect but they "tentatively conclude . . . that the procedures should be expected to generate roughly the same estimates of risk attitudes for a target population . . . when used at the beginning of a session".

By contrast, Deck et al. (2008) compare coefficients of CRRA across three tasks: H&L, E&G and a task which represents a series of binary choices between a risky lottery and an amount of money for certain. They observe significant differences in the

obtained coefficients between tasks and account for these inconsistencies using participants' personality traits. Our study extends the range of comparisons for the E&G and H&L procedures and allows further comparisons with the allocation procedure.

5.2 Choice List and Ranking Procedures

Up to this point we have tried to avoid using any particular functional forms because any such usage is open to the objection that the selected form may be the wrong one – at least, for some, and perhaps many, members of a sample. On the other hand, the instruments discussed in previous sections often *are* used in conjunction with particular specifications.[19] So it may be of interest to consider an example in relation to the H&L and E&G tasks. Were we to assume CRRA, using a standard format whereby $u(x) = x^{1-r}/(1-r)$ for $r \neq 1$ and $u(x) = \ln(x)$ for $r = 1$, we could use the median[20] estimate of r derived from people's responses to the five choice list tasks to predict their top-ranked options in the DT14 ranking task, and then compare those predictions with the actual responses.

Table 10 does that. For both subsamples, there are very clear differences between the actual distributions and those based on the choice list procedure. Most strikingly, the choice list tasks *never* produce a median r high enough to entail ranking the sure £10 option first and only seldom produce a median r greater than 1, whereas the majority of individuals behave in the ranking task as if their values of r are greater than 1.

Table 10: Actual vs Inferred Top Ranked Options in DT14

	V1 Actual DT14	V1 Inferred DT14		V2 Actual DT14	V2 Inferred DT14
A: £10 for sure	70	0		77	0
B: £15, 0.5; £8, 0.5	28	5		22	9
C: £20, 0.5; £6, 0.5	23	25		26	26
D: £25, 0.5; £4, 0.5	13	62		5	54
E: £30, 0.5; £2, 0.5	2	39		4	35
F: £35, 0.5; £0, 0.5	36	40		45	55

So although there may be a significantly positive rank correlation between the two procedures – for example, the Spearman correlation coefficients for DT14 and DT20 are 0.392 for V1 and 0.463 for V2, both of which are significant at the 1% level – the degree of transferability of more precise parameter estimates may be very much more limited. Of course, we acknowledge that part of the reason for this could be

[19] For those readers interested in further analysis on the basis of the most common assumption – that people can be reasonably well modelled as CRRA EU maximisers – we have included a number of such analyses in Appendix D.

[20] We take the median because it is less susceptible to extreme (and sometimes simply confused or mistaken) responses, especially when we have only five observations.

that CRRA is not the appropriate specification for all individuals; but that only serves to underline the fact that the validity of any parameter estimates depends upon the adequacy of the particular assumption made about a functional form.

5.3 The Allocation and Choice List Procedures

In Section 4 we saw that dividing the £20 total in strict proportion to observed probabilities was a popular response, with 55 participants doing so in every one of the thirteen DTs. Even for those who sometimes divided the total in some other way, proportionality was often observed, and overall 150 participants' median response was strictly proportional.

If that median were reasonably diagnostic, we should expect these 150 to behave quite similarly to one another in the choice list decision tasks – that is, to be switching at much the same point as each other in each list. But that is not what we see: although they may be a *little* more tightly clustered than the rest of the sample in DT17 and DT19, the difference is slight. For the other three choice lists, there is no discernible difference: they are typically distributed across four or five switching points in much the same way as those for whom proportional allocation was not the median. The allocation procedure is at best a weak indicator of behaviour in the choice list tasks.

5.4 The Allocation and Ranking Procedures

In this case, we can make direct comparisons between two tasks – DT4 and V2's DT15 – which effectively asked the same question framed differently. In DT4, respondents were asked to divide £20 between Black and White when there were 7 Black and 3 White balls. In the DT15 presented to V2, there were 7 Black and 3 White balls and six discrete options offering different ways of dividing £20 between Black and White. In short, this version of DT15 is essentially the same as DT4 except that the options are restricted to the six involving payoffs which are multiples of £2.

If we were to assume that, when presented with DT15, each respondent would select the option closest to whatever answer she gave in DT4 (with those giving responses exactly halfway between options being assigned in equal numbers to both), we can again compare the distribution of actual responses to DT15 with the distribution inferred from DT4, as in Table 11 below.

It is clear that the actual pattern of responses was very different from the one that would be inferred from the allocation task, with the ranking task pulling towards the less risk averse options E and F while the allocation task pulled the distribution strongly towards the proportional division. Thus nearly three times as many participants top-ranked E and F as would have been predicted from the allocation responses, while about 50% more participants were predicted to top-rank the £14:£6 option C than actually did so – and this, even after they had previously undertaken 13 allocation tasks that might have primed them towards proportionality. Such a difference would be compatible with the general proposition that people are imprecise about their preferences and thus are liable to process what is formally the same problem in rather different ways when the framing provides rather different procedural 'cues'. It is plainly incompatible with any

'deterministic-core-plus-white-noise' model that would entail only random differences between the two distributions.

Table 11: Actual vs Inferred Top Ranked Options in DT15, V2

	Actual DT15	Inferred from DT4	
A:£10 for sure	19	13	(13 + 0)
B:£12, 0.7; £8, 0.3	15	22.5	(19 + 0 + 3.5)
C:£14, 0.7; £6, 0.3	67	92.5	(71 + 3.5 + 18)
D:£16, 0.7; £4, 0.3	22	30.5	(11 + 18 + 1.5)
E:£18, 0.7; £2, 0.3	16	3.5	(2 + 1.5 + 0)
F:£20, 0.7; £0, 0.3	40	17	(17 + 0)

A similar story – indeed, arguably an even more striking story – comes from examining the behaviour of the 55 individuals who divided the £20 in all thirteen allocation tasks *exactly* in proportion to the probabilities. Such unerring consistency might seem to be a strong expression of precise preferences, in which case this subset of respondents might be expected to display high levels of consistency in the next DT they encountered – the ranking task DT14, which was common to both V1 and V2. If we took those individuals' preferences to be logarithmic in gains, there would be a precise prediction for DT14: the optimal option, were it to be available, would be (£17.50, 0.5; £7, 0.5).

In fact, the two closest options that were actually available on either side of that 'optimum' were (£15, 0.5; £8, 0.5). and (£20, 0.5; £6, 0.5). However, only 12 (21.8%) of the 55 individuals put one or other of those options first, while 39 (70.9%) of them chose one of the extreme options – either £10 for sure or (£35, 0.5; £0, 0.5). It might be that, far from the allocation task tapping into preferences with much greater precision, those who stuck most rigidly to a proportionality heuristic were if anything *less* confident about their preferences and were therefore more content to use a simple rule of thumb.

6. Concluding Remarks

It would be convenient if it were true that most people have an 'attitude to risk' which is expressed in a stable and consistent manner across multiple contexts and which is reasonably easy to measure by some simple task. If this *were* true, it would be highly desirable to include a couple of such tasks in experiments or surveys in order to be able to take account of risk attitude when analysing data and interpreting their significance.

However, our results caution against supposing that the three procedures used in a number of recent studies can deliver the desired level of reliability and transferability. In arriving at our conclusions, we have for the most part tried to avoid

assuming any particular functional forms but have relied instead on nonparametric measures that provide much more general tests. The picture emerging from our experiment may be summarised as follows:

1. Except in cases where individuals follow an available rule of thumb, most individuals' responses to different questions within a particular procedure exhibit a degree of variability which appears to increase as the structure and/or parameters of the questions become more dissimilar. Thus it may be unsafe to expect that just one or two questions of *any* kind can provide a reliable measure at the individual level. Moreover, the way in which variability between responses changes as questions become more dissimilar leads us to conjecture that even if one were arguing for a 'deterministic-core-plus-white-noise' model within a procedure, an off-the-shelf standard error specification is unlikely to be adequate: at the very least, we should investigate the extent to which the noise is contextual and/or heteroscedastic (see Buschena and Zilberman, 2000, and Wilcox, 2011).
2. However, this will not suffice. Even when we hold constant the parameters and the type of task, significant differences in patterns of response can still be induced by something as seemingly innocuous as which way up a table is presented. This is consistent with decision making having a potentially influential procedural component.

3. The fact that some individuals display high degrees of consistency in a particular type of task does not necessarily mean either that they have highly articulated underlying preferences or that the task is particularly good at detecting preferences which will transfer to other contexts. In fact, the opposite might be the case: it may be that a number of people who are quite uncertain about their preferences may find it appealing to use a simple heuristic that 'solves' the problem for them. However, this may have little or no predictive power in other tasks where that heuristic is not so readily available.

The overall picture, then, is that most individuals exhibit a good deal of variability in their responses to questions intended to elicit their risk attitudes. There is *some* rank correlation between risk attitudes elicited by different questions, but the imprecision of most people's preferences may make them susceptible to considerable procedural effects.

How should we react to these findings? In the short run, one recommendation is that researchers who wish to take some account of and/or make some adjustment for risk attitude in their studies should take care to pick an elicitation procedure as similar as possible to the type of decision they are studying; and ideally, should use several different questions and/or at least two different procedures in order to check the sensitivity of the risk attitude parameter estimates they generate.

In the longer run, the challenge is to engage with the inherently stochastic nature of human decision making and develop models of the *processes* which produce people's responses. Deterministic models may be analytically more tractable, but

they are not realistic; and adding some more or less arbitrary random error term to a deterministic core will not make them so. If the variability in human judgment is a reflection of decision making as a cognitive process, we need to try to gain a better understanding of how contextual or procedural factors interact with that process. Wishing such influences away and assuming that decision processes are reducible to one-size-fits-all sets of axioms has not and will not produce a descriptively adequate account of human behaviour under risk and uncertainty.

References

Andreoni, J. and C. Sprenger (2012a) "Estimating Time Preferences From Convex Budgets." *American Economic Review*, *102*(7), pp. 3333-3356.

Andreoni, J. and C. Sprenger (2012b) "Risk Preferences Are Not Time Preferences." *American Economic Review*, *102*(7), pp. 3357-3376.

Andersen, S., Harrison, G., Lau, M. and Rutstrom, E. (2008) "Eliciting Risk and Time Preferences" (with Steffen Andersen, Morten Lau and Elisabet Rutström), *Econometrica*, *76*, 583-618.

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C. and R. Sugden (2009) *Experimental Economics: Rethinking the Rules*, Princeton, NJ: Princeton University Press.

Binswanger, H. P. (1980) "Attitudes toward Risk: Experimental Measurement in Rural India." *American Journal of Agricultural Economics*, *62*(3), pp. 395-407.

Binswanger, H. P. (1981). "Attitudes toward Risk: Theoretical Implications of an Experiment in Rural India." *Economic Journal*, *91*(364), pp. 867-890.

Blavatskyy, P. R., and G. Pogrebna (2010) "Models of Stochastic Choice and Decision Theories: Why both Are Important for Analyzing Decisions." *Journal of Applied Econometrics*, *25*(6), pp. 963-986.

Brown, A. L., and H. Kim (2013) "Do Individuals Have Preferences Used in Macro-Finance Models? An Experimental Investigation." *Management Science*, forthcoming.

Busemeyer, J. and Townsend, J. "Decision Field Theory: A Dynamic-cognitive Approach to Decision Making in an Uncertain Environment." *Psychological Review*, *100*, 432-459

Buschena, D. and D. Zilberman (2000) "Generalized Expected Utility, Heteroscedastic Error and Path Dependency in Risky Choice," Journal of Risk and Uncertainty, 20(1), pp. 67-88. [Note erratum, (2008), 36(2), p.201]

Butler, D. and G. Loomes (2007) "Imprecision as an Account of the Preference Reversal Phenomenon." *American Economic Review*, *97*, pp. 277-297.

Charness, G., & Gneezy, U. (2012). "Strong Evidence for Gender Differences in Risk Taking." *Journal of Economic Behavior & Organization*, *83*(1), 50-58.

Cheung, S. (2013). "On The Elicitation of Time Preference Under Conditions of Risk," University of Sydney Working Paper #2013-15.

Choi, S., Fisman, R., Gale, D. and S. Kariv (2007) "Consistency and Heterogeneity of Individual Behavior under Uncertainty," *American Economic Review*, *97*(5), pp. 1921-1938.

Cohen, M., Jaffray, J. Y., and T. Said (1987) "Experimental Comparison of Individual Behavior under Risk and under Uncertainty for Gains and for Losses." *Organizational Behavior and Human Decision Processes*, *39*(1), pp. 1-22.

Crosetto, P. and A. Filippin (2013) "A Thorough Investigation of Gender Differences in Risk Attitudes Using the Holt and Laury Elicitation Task," mimeo.

Cubitt, R., Navarro-Martinez, D. and C. Starmer (2013) "On Preference Imprecision," mimeo.

Dave, C., Eckel, C., Johnson, C. A. and C. Rojas (2010) "Eliciting Risk Preferences: When Is Simple Better?" *Journal of Risk and Uncertainty*, *41*, pp. 219-243.

Deck, C., Lee, J., Reyes, J. and C. Rosen (2008) "Measuring Risk Attitudes Controlling for Personality Traits," *mimeo*, available at **http://www.econ.canterbury.ac.nz/research/pdf/Paper_Deck.pdf**

Dubourg, R., Jones-Lee, M. and G. Loomes (1997) "Imprecise Preferences and Survey Design in Contingent Valuation," *Economica*, *64,* pp. 681-702.
Eckel, C. and P. J. Grossman (2002) "Sex Differences and Statistical Stereotyping in Attitudes toward Financial Risk." *Evolution and Human Behavior*, 23, pp. 281-295.
Harrison, G. and E. Rutstrom (2008) "Risk Aversion in the Laboratory," in J.Cox and G.Harrison (eds), *Risk Aversion in Experiments*, Research in Experimental Economics, Volume 12, (Bingley, UK: Emerald).
Hey, J. and N. Pace (2011) "The Explanatory and Predictive Power of Non Two-Stage-Probability Theories of Decision Making Under Ambiguity," Working Paper, Department of Economics, University of Venice "Ca' Foscari", accessed at http://ideas.repec.org/p/ven/wpaper/2011_12.html
Holt, C. and S. Laury (2002) "Risk Aversion and Incentive Effects," *American Economic Review*, *92*(5), pp. 1644-1655.
Jamieson, D. and Petrusic, W. (1977) "Preference and Time to Choose," *Organizational Behavior and Human Performance*, 19, pp. 56-67.
Kahneman, D. and A. Tversky (1979) "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2), pp. 263-291.
Kahneman, D. and A. Tversky (Eds.). (2000) *Choices, Values, and Frames*. Cambridge University Press.
Lévy-Garboua, L., Maafi, H., Masclet, D. and A. Terracol (2012) "Risk Aversion and Framing Effects," *Experimental Economics*, *15*(1), pp. 128-144.
Loomes, G. (1991) "Evidence of a New Violation of the Independence Axiom," *Journal of Risk and Uncertainty*, *4*(1), pp. 377-400.
Moffatt, P. (2005) "Stochastic Choice and the Allocation of Cognitive Effort." *Experimental Economics, 8*, pp. 369-388.
Mosteller, F. and P. Nogee (1951) "An Experimental Measurement of Utility." *Journal of Political Economy*, *59*, pp. 371-404.
Otter, T., Johnson, J., Rieskamp, J., Allenby, G.M., Brazell, J.D., Diedrich, A., Hutchinson, J.W., MacEachern, S., Ruan, S. and J. Townsend (2008) "Sequential Sampling Models of Choice: Some Recent Advances", *Marketing Letters*, *19*, pp. 255-267.
Parducci, A. (1965) "Category Judgment: a Range-Frequency Model." *Psychological Review*, *72*(6), pp. 407-18.
Rabin, M. (2000) "Risk Aversion and Expected Utility Theory: A Calibration Theorem," *Econometrica*, *68*(5), pp. 1281-1292.
Simon, H. A. (1978) "Rationality as Process and Product of Thought." *American Economic Review*, *68(2)*, pp. 1-16.
Stott, H. P. (2006) "Cumulative Prospect Theory's Functional Menagerie." *Journal of Risk and Uncertainty*, *32*(2), pp. 101-130.
Tversky, A. and D. Kahneman (1992) "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty*, *5*(4), pp. 297-323.
Wilcox, N. (2011) "Stochastically More Risk Averse: A Contextual Theory of Stochastic Discrete Choice Under Risk." *Journal of Econometrics*, 162(1), pp.89-104.

Appendix A: Experimental Procedure and Instructions

The experiment was computerised using Qualtrics Software. We invited 800 undergraduate students at the University of Warwick via the electronic volunteers' register of the Decision Research at Warwick (DR@W) group. Students were invited to participate in the experiment online at any time during a five-day period (from 22.11.2010 to 26.11.2010) on the understanding that 50 of those who participated would be picked at random to come to the laboratory at a mutually convenient time and enact one of their decisions – also randomly selected – and be paid according to how that decision played out.[21] We took various precautions to ensure that no individual participated more than once.[22]

Of the 800 invited to take part, 400 randomly selected people were sent a link to the 'treatment' variation 1 (V1) of the experiment and the other 400 received a link to the variation 2 (V2). Both variations of the experiment were designed using the Qualtrics software. Experimental tasks were accessible via the following links on the Qualtrics website[23]:

- V1: https://columbia.qualtrics.com/SE/?SID=SV_9GH9H9WSptIYgJK
- V2: https://columbia.qualtrics.com/SE/?SID=SV_1Mte22jrFxXKYJu

People were asked to follow the link and read the experimental instructions. On the first screen of the experiment, potential participants were told that the experiment consisted of 20 questions. They were also informed that all participants who complete the experiment by the specified deadline would enter a random draw. 50 participants, selected in this random draw, would then be invited to the experimental laboratory and would have an opportunity to play out their decisions (made online) for real money.

By the end of the five-day period, 423 completed sets of decisions had been submitted: 206 for V1 and 217 for V2. We recorded the time participants took to complete the experiment (this was the time between starting to answer the first task and submitting the last decision in the experiment). The median completion time was 19 minutes across the two versions (median completion times for both versions were very similar: 18 minutes for V1 and 19 minutes for V2). 319 participants (75%) took less than 30 minutes to complete the experiment. Even though participants had an opportunity to complete the experiment over the course of 5 days (dropping and picking up experimental tasks at any point), only 7 people (2%) took more than 1 day to complete all the tasks.

To ensure transparency, the ID numbers of the 50 people picked at random to be paid were reported to all 423 participants (the participants received information about the total number of replies received by the experimental team and, therefore,

[21] Several recent papers on choice under risk and uncertainty show that incentive mechanisms where each participant is paid according to her decisions and where several participants are selected at random and paid for their decisions produce very similar results (Abdellaoui et al., 2011; Trautmann et al., 2011).

[22] Each participant had previously signed up on the DR@W register of volunteers and was personally invited to the experiment and received a unique sign-in ID. Only one reply from each ID was allowed.

[23] Both versions of the experiment are currently accessible via the links specified above.

had complete knowledge of the odds to be selected). An additional e-mail was sent to the randomly selected participants explaining when and how they can play out their decisions and receive their payoffs. These participants were given one working week (following the experiment) to come to the experimental laboratory and play out their decisions made online. 41 out of 50 randomly selected participants tuned up during the allocated week, 8 made special arrangements and were paid after they came back to the University from their Christmas holidays. One person who was randomly selected has never turned up and never made contact with the experimental team despite receiving multiple reminders. The average payoff in the experiment was £10.90.[24]

The screenshot of the first screen is displayed below.

We are inviting you, as someone who signed up with **DR@W**, to take part in an on-line experiment. To participate, you will need to make 20 decisions, which we think will take you about 20 -30 minutes, including the time taken to read and understand the instructions. You could be paid up to £35 on the basis of one of your decisions – although it is also possible you may get zero. That is the nature of decision making when some degree of risk or uncertainty is involved!

The decisions come in three groups. First, a set of 13 decisions where you are asked to divide money up between different things. Second, 2 decisions where we ask you to rank six options involving certain chances of different sums of money. Third, a set of 5 decisions involving choices between various pairs of alternatives.

There is no single 'right' answer in any one of these decisions – different people may have different preferences and we simply want you to tell us YOUR personal preference.

When you have made all of your decisions and have submitted them to us, here is what will happen. We will look at all those who have submitted their decisions by

NO LATER THAN 16.00 (that is, 4 p.m.) on Friday 26th November 2010

and then we will select at random 50 of those people and invite them to come to the DR@W laboratory and play out one of their decisions for real. The one to be played for real is picked at random and each person's payment will depend ENTIRELY on how that one randomly-selected decision works out. So it is in your interests to think carefully about each decision in turn and only submit your answers once you are really happy with them all.

Once you have submitted your decisions, you will NOT later be allowed to change any of them.

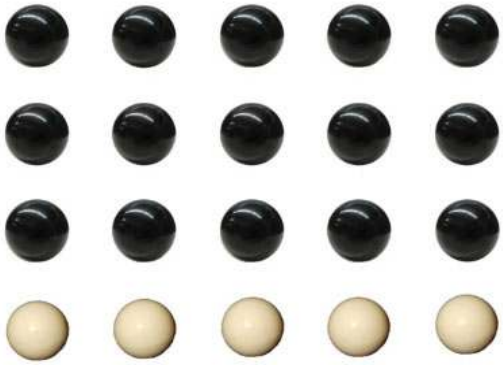
After reading the general instructions, participants received specific instructions to the allocation procedure which consisted of 13 decision tasks. There instructions are displayed in the next page.

[24] Further details about the experimental procedure are given in Appendix A.

FIRST 13 DECISIONS

The General Idea

Imagine a bag containing 20 coloured balls **– 15 black** and **5 white.**
Suppose that the bag is going to be shaken to mix the balls up and then one ball will be picked at random.



BEFORE the ball is picked, you are given £20 to divide any way you like between black and white. You will be paid whatever amount you allocate to the colour of the ball that is picked. All you know at the time you split the money is that there is a 75% chance the ball will be black and a 25% chance it will be white.

How would you divide the £20 between black and white?

From past experience, we expect different people to do different things. Some might like to split the money £10 : £10 to be sure of getting £10 whichever colour is drawn. Others might put all £20 on black and 0 on white to give themselves a 75% chance of receiving £20 but also a 25% chance of getting nothing. Others may put at least some money on white but put more on black. Thinking about the various ways you could divide the £20, which would YOU go for?

That was an example which was just to explain the idea.

We would now like you to make a number of similar decisions for real money. Unlike the example, these all involve TEN balls and up to THREE colours.

On the next sheet you will see 13 different ways that 10 balls can be divided between two or three colours (besides black and white, some bags will contain some yellow balls). In each case, we would like you to allocated £20 between them (a fresh £20 each time) in the way that you personally would like best. If you prefer to put all £20 on a single colour, you are free to do so. If you want to spread the £20 as equally as possible between three colours, you can do that – the amounts must be multiples of 10p, so that when there are three colours you can divide the £20 up as £6.70 : £6.70 : £6.60 if you wish. Or you can decide on any other ways of allocating the £20 that you like better. Give it some careful thought, then fill in the spaces on the next page, and when you are happy with all your decisions, move on to the next group of 2 decisions.

These specific instructions were followed by 13 decision tasks (DT1-DT13) displayed on the same screen. These decision tasks were the same in both versions of the experiment (V1 and V2). Participants could scroll up and down using the computer mouse to make choices as well as to change their previous choices before they proceeded to the ranking procedure. In the ranking procedure, participants received specific instructions (displayed below) followed by two decision tasks (DT14 and DT15). DT14 was the same for both versions of the experiment and DT15 differed between the two versions.

NEXT 2 DECISIONS

In each of these two decisions there are six options. Which of the six do you like most? Put a 1 in the box next to that option. Now look at the other five: which of those is your 'next-best' choice? Put a 2 in the box next to that one. Put a 3 against your third-best – and so on, down to 6 against the one that would be your least preferred of these six options.

If one of these questions is played out for real, we will pick two of the options at random and you will play out whichever one of the two you have ranked higher.

Finally, after completing the ranking procedure, participants proceeded to the choice list procedure. At that point, they saw specific instructions followed by the last 5 decision tasks (DT16-DT20). These specific instructions are displayed below:

FINAL 5 DECISIONS

Each of these decisions involves comparing uncertain options with some sure sums of money. In each case, one option will stay fixed while the other gets progressively better or worse. You are asked for each comparison to say which you prefer by ticking the boxes. So in Decision 16, below, in the first row you are asked to decide whether you would rather have a 60% chance of £10 plus a 40% chance of getting 0 (drawing from a bag with 6 black balls and 4 white balls, where black gives £10 and white gives 0) or else a sure £10. Once you have clicked on the tick box next to the option you prefer, go to the next row. Now you are asked to choose between the 60% chance of £10 plus a 40% chance of 0 or else £9 for sure.

Continue to work your way down each row, ticking whichever box on that row indicates the option you would prefer to have. Please take care to enter one tick per row for all 10 rows. Then go to the next decision and again indicate your preference for each of the 10 rows there.

If one of the next 5 decisions is to be played out for real, we will pick one of the 10 rows at random and your payment will be determined by whatever you ticked on that row. If you have ticked the sure amount, you will simply receive that sum. If you have ticked the uncertain option, we will fill a bag with the appropriate mix of black and white balls and your payment will depend on which ball is picked out of the bag at random.

Each of the 50 participants selected to play out one of their decisions was offered an individual 5-minute appointment. When they came, they selected one decision task at random and were paid according to their decision in that task. If in a particular decision task, a lottery was selected, a participant would receive an outcome of that lottery. If a sure amount was selected, a participant would receive that sure amount. Payments were made in cash at the end of each individual appointment.

Appendix B: Exclusion Criteria

In Appendix B we discuss how far responses in our 3 elicitation procedures conform with principles typically assumed in the analysis of decision making (such as respect for dominance). We identify grounds for setting aside some of our 423 participants' responses from the main analysis.

As noted in Section 1 of the manuscript, most people's responses to many experimental decision tasks exhibit some stochastic component. Some variability of this kind appears to be intrinsic to brain function and it would therefore be too restrictive to exclude participants on the grounds that they exhibit any inconsistency at all. On the other hand, most people who have run experiments will acknowledge that there are often a (hopefully small) minority of participants who fail to understand the tasks and/or who adopt rather eccentric response strategies. It may not always be straightforward to draw a sharp line between random error and deeper misconceptions, but in order to reduce the risk that overall patterns are driven too much by confused outliers, we sought to try to set to one side any participants who appeared to exhibit more systematic signs of confusion (although in Appendix C we repeat the main forms of analysis for the full sample, in order to show how far our exclusion policy changes any conclusions we draw).

(a) The Allocation Procedure

In each of DT1, DT2, DT4 and DT6 there were only two colours, Black and White, with more Black balls than White balls. We should therefore expect respect for stochastic dominance to require that at least as much money is allocated to Black as to White in each case. The use of the slider might have meant that someone in haste could have allocated a little more to White than to Black even when they intended to allocate equal sums to both (reflecting extreme risk aversion); but if the amount allocated to White was 50p or more greater than the amount allocated to Black[25], we counted this as an inconsistency.

In fact, there were very few such inconsistencies. One individual adopted an 'inverse matching' strategy, dividing the money in inverse proportion to the probabilities – i.e., dividing the £20 2:18 in DT1, 4:16 in DT2, 6:14 in DT4 and 8:12 in DT6. There were ***no*** other inconsistencies among the other 422 participants in DT2 or DT4, one other in DT1 and two others in DT6. In these most basic allocation tasks, then, more than 99% of responses satisfied a minimal consistency condition.

In DT3, DT8 and DT13, the numbers of White and Yellow balls were the same. We should therefore expect the amounts allocated to be the same (except for small differences that might be attributed to hasty use of the slider). In DT12, there were four Black, four White and two Yellow balls, so that we should expect the amounts allocated to Black and White to be the same (except for small 'slider errors' and in cases where an individual behaves as if risk neutral, allocating zero to Yellow and obtaining equal expected value with any division between Black and White).

[25] It could be argued that 50p is an arbitrary cut-off, and to some extent it is, but the dataset suggests it is the smallest 'round' amount that people in this task used to signify an intentional difference.

The numbers of cases where the 'equality' requirement was breached by a difference of 50p or more varied across these four questions: for DT3 it was breached by 11 participants; for DT8, by 25; for DT12 by 21; and for DT13, by 39. The stronger form of inconsistency, where an individual allocated at least 50p more to a colour with a strictly lower probability of occurring, was very much less common: apart from the individual who followed the 'inverse proportion' strategy rigorously throughout, there was just one instance for DT3, two for DT8, two for DT12 and eight for DT13.

In DT5, DT7, DT10 and DT11, Black was always strictly more probable than White, which in turn was always strictly more probable than Yellow. So inconsistencies in these questions involved allocating at least 50p more to Yellow than to White or Black and/or allocating at least 50p more to White than to Black. Apart from the individual using the 'inverse proportion' strategy throughout, there were, respectively, three, two, four and five such inconsistencies.

So it appears that when one probability is strictly greater than another, there are few violations of dominance, suggesting that the great majority of participants paid attention to the stimuli and satisfied a basic requirement of rationality. When two colours were equiprobable, *some* participants found it less compelling that they should allocate the same amount to each, although the clear majority did go for an equal allocation ('slider error' aside). If a single breach of the 'equality' requirement was the only form of inconsistency, we did not exclude the participant's data from the analysis. In the allocation procedure, therefore, we excluded from the analysis the 12 participants who violated dominance by 50p or more and a further 19 who breached the equality requirement by more than 50p on at least two occasions.

(b) The Ranking Procedure

Although it is usually assumed that preferences are single-peaked, this is an assumption that does not appear to have been much examined. A comprehensive review of all possible rankings and their compatibility with single-peakedness would require a long and complex discussion, but some basic analysis may be informative. An individual whose preferences are single-peaked could be expected to identify his second-ranked option as being adjacent to his top-ranked option; and his third-ranked option would then be expected either to be adjacent to the second-ranked option or else adjacent to the first-ranked option on the other side. Confining our present analysis to just the first three ranked items, Table B1 reports the results for DT14 (which was presented to the whole sample) and DT15 (which was different for the two versions V1 and V2).

The row labelled 'First three consistent' shows how many participants gave rankings 1, 2 and 3 in a pattern consistent with single-peakedness as described above. The row labelled 'First two only adjacent' reports the numbers of participants whose second-ranked option was adjacent to the first-ranked option but whose third-ranked option was not adjacent to either of the first two. The row labelled 'First and second not adjacent' records how many participants did not fulfil this most basic requirement of single-peakedness.

On average, the first three rankings were consistent with single-peakedness in just over 70% of responses, while 12.3% of the total did not satisfy adjacency for the two highest ranked options. Of this latter group, the most prevalent 'deviation' involved the first two rankings being at opposite ends of the table: out of the total of 104 observations in the bottom row of Table B1, 68 exhibited this pattern.

Table B1: Consistency with Single-peakedness

	DT14		DT15	
	V1	**V2**	**V1**	**V2**
First three consistent	131 (63.6%)	144 (66.4%)	167 (81.1%)	154 (71.0%)
First two only adjacent	42 (20.4%)	51 (23.5%)	17 (8.3%)	36 (16.6%)
First and second not adjacent	33 (16.0%)	22 (10.1%)	22 (10.7%)	27 (12.4%)

A single case where the first and second ranked options are not adjacent might be attributable to error, but if we observe such non-adjacency in *both* of an individual's responses to these questions, we omit that individual from the main analysis: 25 participants came into this category, of whom two were already excluded on the basis of their allocation procedure responses, bringing the total of exclusions on the basis of the allocation and ranking procedures up to 54.

(c) The Choice List Procedure

The design of this procedure implies that participants with standard preferences should make at most one switch from the left hand side of a particular table to the right hand side, or vice-versa. For each task of this type (DT16-DT20) we can identify cases when a participant switched more than once. If this pattern could be regarded as the result of a single misjudged choice – that is, if by changing just one choice in one row, we could produce a single switching point – we classify this as a 'weak' inconsistency. On the other hand, if we observe multiple switching and/or a pattern of choice which violates dominance in a way that cannot be resolved by a single change, we classify this as a 'strong' inconsistency.

Across the five choice list tasks, 379 participants exhibited no inconsistencies at all, and a further 28 exhibited just one weak inconsistency, which might be attributed to a momentary lapse of attention. 26 other participants exhibited at least one strong inconsistency and/or more than one weak inconsistency, and we have excluded them from the analysis. Allowing for the fact that some of these were also excluded according to criteria from the allocation procedure and/or the ranking procedure, this brings the total number of exclusions to 72, leaving 351 (172 in V1 and 179 in V2) who form the basis of the analysis presented in the paper.

Appendix C: Results for All Participants

In Appendix B we explained that 72 participants were excluded from consideration in our analysis due to various reasons (e.g., violation of dominance). In Appendix C we conduct a robustness check and show that results which are reported in the main body of the paper do not change if we include all 423 participants' decisions into our dataset.

Table C1 replicates Table 1 for the entire sample.

Table C1: Distributions of Responses in DT16–DT20

Sure/Safer	DT16 (£10, 0.6; 0, 0.4) vs £10 to £1		DT17 (£18, p; £3, 1-p) vs £8		DT18 (£18, 0.2; £3, 0.8) vs £10 to £1		DT19 (£15, p; 0, 1-p) vs £5		DT20 (£8, p; £5, 1-p) vs (£20, p; £1, 1-p)	
	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2
0	1	3	1	1	1	1	1	-	2	1
1	3	1	1	1	-	1	-	2	3	1
2	2	2	6	8	4	11	6	6	26	26
3	19	5	45	63	8	23	46	47	50	26
4	49	42	69	70	39	53	62	73	56	52
5	58	66	44	45	71	70	38	44	35	58
6	50	65	28	20	49	31	42	20	25	34
7	17	26	9	7	28	23	8	15	8	14
8	3	4	1	2	1	3	1	7	-	3
9	-	2	2	-	1	-	2	3	1	2
10	4	1	-	-	4	1	-	-	-	-
Average Sure/Safer choices	**5.01**	**5.30**	**4.37**	**4.13**	**5.29**	**4.78**	**4.48**	**4.49**	**3.98**	**4.48**

Tables C2 and C3 show the rank correlation coefficients for each pair of tasks within each subsample.

Table C2: Spearman Rank Correlation Coefficients, Choice List Procedure, V1

	DT17	**DT18**	**DT19**	**DT20**
DT16	0.150*	0.224**	0.331***	0.404***
DT17		0.276***	0.626***	0.400***
DT18			0.286***	0.247**
DT19				0.538***

* significant at 0.05; ** significant at 0.01; *** significant at 0.001 level

Table C3: Spearman Rank Correlation Coefficients, Choice List Procedure, V2

	DT17	**DT18**	**DT19**	**DT20**
DT16	0.241**	0.172*	0.293***	0.415***
DT17		0.225**	0.532***	0.393***
DT18			0.183**	0.155*
DT19				0.475***

* significant at 0.05; ** significant at 0.01; *** significant at 0. 001 level

In all cases, the correlations are positive and statistically significant, and in both subsamples they are highest for the two most similar tasks, DT17 and DT19. This confirms our results reported for the fraction of the participants in the main body of the paper that there is at least some degree of heterogeneity between individuals, with some tending to be more or less risk averse than others, but for comparisons other than DT17-DT19 the coefficients are low, which is compatible with a good deal of imprecision and susceptibility to procedural effects.

Table 4 in the main paper was a table of parameters. Tables C5 and C6 replicate Tables 5 and 6 and provide a summary of the comparison between DT14 and DT15 in the ranking procedure.

Table C5: 1st Choices in DT14 and DT15, V1 subsample

DT14 1st Choice		**DT15 1st Choice**	
A: £10 for sure	86	134	**A: £10 for sure**
B: £15, 0.5; £8, 0.5	37	29	**B: £15, 0.3; £8, 0.7**
C: £20, 0.5; £6, 0.5	25	14	**C: £20, 0.3; £6, 0.7**
D: £25, 0.5; £4, 0.5	14	4	**D: £25, 0.3; £4, 0.7**
E: £30, 0.5; £2, 0.5	2	3	**E: £30, 0.3; £2, 0.7**
F: £35, 0.5; £0, 0.5	42	22	**F: £35, 0.3; £0, 0.7**

Table C6: 1st Choices in DT14 and DT15, V2 subsample

DT14 1st Choice		**DT15 1st Choice**	
A: £10 for sure	92	27	**A: £10 for sure**
B: £15, 0.5; £8, 0.5	29	17	**B: £12, 0.7; £8, 0.3**
C: £20, 0.5; £6, 0.5	31	74	**C: £14, 0.7; £6, 0.3**
D: £25, 0.5; £4, 0.5	6	29	**D: £16, 0.7; £4, 0.3**
E: £30, 0.5; £2, 0.5	5	16	**E: £18, 0.7; £2, 0.3**
F: £35, 0.5; £0, 0.5	54	54	**F: £20, 0.7; £0, 0.3**

Results presented in Tables C5 and C6 are remarkably similar to the results summarized in Tables 5 and 6 in the main body of the paper. In Table C5, the direction of movement is as expected, although of the 86 who chose the sure option in DT14, 17 moved away to a riskier position in DT15, despite those options offering much poorer gains in return for the extra risk. At the same time, the 92 who opted for A in DT14 (and therefore turned down the mean-variance trade-off in B or beyond) should not on that basis have opted for anything riskier than B in DT15 (see Table C6). However, only 29 participants select lotteries according to this prediction in DT15, while the other 63 opt for much riskier alternatives.

Table 7 in the main paper showed the parameters for the allocation tasks. Table C8 below categorises responses to the tasks that involved only Black and White balls in ratios greater than 1.

Table C8: Allocation to Black in DT1, DT2, DT4 and DT6

£ in Black	DT1 0.9 : 0.1	DT2 0.8 : 0.2	DT4 0.7 : 0.3	DT6 0.6 : 0.4
<10	2	1	1	3
10	6	7	17	88
	-	1	5	17
11	-	-	1	33
	1	1	3	8
12	1	2	34	173
	-	3	14	23
13	2	5	27	14
	1	2	10	7
14	-	4	128	8
	2	2	13	-
15	33	70	83	22
	4	16	13	4
16	7	117	13	4
	1	9	2	-
17	19	20	11	1
	13	12	3	-
18	106	36	6	-
	8	4	1	-
19	23	10	-	-
	7	4	-	-
20	187	97	38	18

Table C9 provides a summary of the Spearman rank correlations in the allocation task for all 423 participants.

Table C9: Spearman Rank Correlation Coefficients,

Allocation Procedure for (all) 423 Participants

	DT2	DT3	DT4	DT5	DT6	DT7	DT8	DT10	DT11	DT12	DT13
DT1	0.707	0.705	0.511	0.606	0.213	0.513	0.338	0.622	0.381	0.421	0.175
DT2		0.815	0.687	0.694	0.374	0.595	0.484	0.553	0.446	0.509	0.263
DT3			0.634	0.733	0.377	0.633	0.525	0.570	0.460	0.430	0.245
DT4				0.789	0.478	0.575	0.535	0.464	0.445	0.463	0.317
DT5					0.474	0.719	0.581	0.591	0.495	0.538	0.325
DT6						0.619	0.682	0.277	0.390	0.366	0.303
DT7							0.730	0.628	0.546	0.529	0.348
DT8								0.390	0.554	0.499	0.389
DT10									0.542	0.530	0.272
DT11										0.597	0.368
DT12											0.387

All correlation coefficients are statistically significant at the 1% level. However, there is considerable variability in the coefficients, from 0.815 down to 0.175. The correlations tend to be highest between those tasks that are adjacent and/or have the most similar parameters, and tend to fall for pairs that are further apart and/or more different. Generally, the correlation coefficients reported in Table C9 are lower than those in Table 9 in the main paper. But overall, the behavioural patterns reported in the paper for 351 participants are very similar to those for all 423 participants.

Appendix D: Estimations of CRRA Coefficients

(a) Within-procedural comparison

In the main paper, most of our analysis tried to avoid using any specific form of utility function. Yet, our analysis can also be done with a functional form in mind. For example, our results can be replicated if we assume the CRRA utility function where $u(x) = x^{1-r}/(1-r)$ for $r \neq 1$ and $u(x) = \ln(x)$ for $r = 1$.

Table D1 shows the variability across the different decision tasks presented in the Allocation Procedure in terms of the percentages behaving as if $r > 1$, or as if $r = 1$, or as if $r < 1$: the degree of variability is clearly at odds with the idea that each individual has some 'true' risk coefficient r*, with each person's revealed r on any occasion deviating from their r* only randomly.[26] Therefore, values of the coefficient r vary significantly within the Allocation Procedure.

Table D1 Percentages in Different r Categories, by Question

CRRA coefficient r	**DT1** **9:1**	**DT2** **8:2**	**DT3** **8:1:1**	**DT4** **7:3**	**DT5** **7:2:1**	**DT6** **6:4**	**DT7** **6:3:1**	**DT8** **6:2:2**	**DT10** **5:4:1**	**DT11** **5:3:2**	**DT12** **4:4:2**	**DT13** **4:3:3**
r < 1	52.1	45.0	43.6	43.9	39.9	22.5	28.8	25.1	21.9	29.6	35.9	30.5
r = 1	27.4	29.6	28.8	31.6	25.6	43.9	31.3	43.3	32.8	43.3	53.8	57.5
r > 1	20.5	26.4	27.6	24.5	34.5	33.6	39.9	31.6	45.3	27.1	10.3	12.0

Results from the Ranking Procedure in V1 are summarised in Table D2. Table D2 captures patterns of the top lottery choices together with the values of r which correspond with indifference between adjacent options. For example, a CRRA individual for whom r = 1 would be indifferent between (£15, 0.5; £8, 0.5) and (£20, 0.5; £6, 0.5) in DT14, while an individual for whom r = 0.6062 would be indifferent between (£20, 0.5; £6, 0.5) and (£25, 0.5; £4, 0.5). So if we find that someone ranks (£20, 0.5; £6, 0.5) as their top choice in DT14, we infer that their r lies somewhere between 1 and 0.6062.

[26] Note that according to the CRRA, an individual is risk seeking if $r<0$, risk neutral if $r=0$ and risk averse if $r>0$. We present the data according to three different categories ($r<1$, $r=1$, $r>1$) because in the Allocation Procedure participants who are risk seeking, risk neutral and slightly risk averse all may divide £20 by constantly betting all of their endowment on black. Therefore, in this case, the lack of stability of r can be better demonstrated by categories $r<1$, $r=1$, and $r>1$ than by categories $r<0$, $r=0$, and $r>0$.

Table D2 Top lottery choices in DT14 and DT15, subsample V1

Value of	DT14 1st Choice		DT15 1st Choice		Value of r
2.958	**A: £10 for sure**	70	118	**A: £10 for sure**	0.2137
1	**B: £15, 0.5; £8, 0.5**	28	22	**B: £15, 0.3; £8, 0.7**	0.0753
0.6062	**C: £20, 0.5; £6, 0.5**	23	9	**C: £20, 0.3; £6, 0.7**	0.0457
0.4089	**D: £25, 0.5; £4, 0.5**	13	2	**D: £25, 0.3; £4, 0.7**	0.0309
0.2334	**E: £30, 0.5; £2, 0.5**	2	2	**E: £30, 0.3; £2, 0.7**	0.0182
	F: £35, 0.5; £0, 0.5	36	19	**F: £35, 0.3; £0, 0.7**	

Given that in DT15 the EV advantage of the riskier options is greatly reduced, we should expect to see more safe choices than in DT14. On the one hand, we do observe more safe choices in DT15. However, it could be argued that the shift in the distribution should have been even stronger between DT14 and DT15. Given their choices in DT14, we may expect at least 136 individuals in V1 have r > 0.2334 and that *all* of these should opt for the sure £10 in DT15. Nevertheless, the shift is significant enough to support the claim that participants are sensitive to the change in parameters and for the most part react in the expected direction.

A similar table can be compiled for V2 (see Table D3 below).

Table D3 Top lottery choices in DT14 and DT15, subsample V2

Value of r	DT14 1st Choice		DT15 1st Choice		Value of r
2.958	**A: £10 for sure**	77	19	**A: £10 for sure**	4.1492
1	**B: £15, 0.5; £8, 0.5**	22	15	**B: £12, 0.7; £8, 0.3**	1.3562
0.6062	**C: £20, 0.5; £6, 0.5**	26	67	**C: £14, 0.7; £6, 0.3**	0.7638
0.4089	**D: £25, 0.5; £4, 0.5**	5	22	**D: £16, 0.7; £4, 0.3**	0.4808
0.2334	**E: £30, 0.5; £2, 0.5**	4	16	**E: £18, 0.7; £2, 0.3**	0.2491
	F: £35, 0.5; £0, 0.5	45	40	**F: £20, 0.7; £0, 0.3**	

As in V1, the distribution of V2 top choices in DT15 is different from the distribution of DT14 choices in the predicted direction. However, when we compare the two distributions, there are grounds for concern about the robustness of the estimates of r. For example, if the 77 individuals who chose the sure £10 option in DT14 all had r > 2.958, we should have expected them either to choose the sure £10 in DT15 or else

choose the next safest option (£12, 0.7; £8, 0.3). But in fact only 21 of those individuals made those choices, while the other 56 spread (i.e., 39, 7, 5, 5) between the other four options. In short, the internal consistency of the ranking procedure based on the E&G question format is modest.

Table D4 shows the distributions of r derived from participants' decisions in the Choice List Procedure. This table reveals significant variability in the distributions of r between decision tasks.

Furthermore, as shown in Table D5, the subsample mean and median estimates of r vary considerably between questions, with the subsample means varying from -0.097 to 0.690 for V1 and from 0.185 to 0.661 for V2, while the subsample medians varied from 0.146 to 0.561 for V1 and from 0.146 to 0.789 for V2. Clearly, such data cast serious doubt on the idea that any one choice list question in conjunction with the CRRA assumption can produce estimates of r that transfer even to other choice list questions with rather different parameters.[27]

What is not so readily apparent from these tables is the degree of variability at the individual level. However, if we compute the ranges and standard deviations of the values of r elicited from each participant, we find that the mean and median ranges were 1.225 and 1.013, while the mean and median within-person standard deviations were 0.491 and 0.420. In short, the data from the choice list tasks tell the same story as the data from the allocation tasks and from the ranking tasks: namely, that many individuals do not exhibit CRRA coefficients that are stable from one set of parameters to another, even within the same type of procedure.

[27] The analysis reported in Appendix D utilises the individual-level data. We have also conducted maximum likelihood estimations of representative agent data using expected utility theory with constant relative risk aversion. Our results at the representative level also show that participants are not consistent between different choice list talks in the experiment. These results are available from the authors upon request.

Table D4 Distributions of r derived from DT16 – DT20

Sure/ Safer	DT16				DT17				DT18				DT19				DT20			
	V1	V2	*r*	*r**	V1	V2	*r*	*r**	V1	V2	*r*	*r**	V1	V2	*r*	*r**	V1	V2	*r*	*r**
0	-	-	-	-	1	-	-2.662	-1.7633	1	1	-1.949	-1.6835	1	-	-1.727	-1.0959	2	1	-1.206	-0.7066
1	3	-	-8.959	-3.8484	1	1	-1.214	-0.8039	-	1	-1.441	-1.2172	-	2	-0.727	-0.4650	3	-	-0.405	-0.1816
2	1	-	-2.143	-1.2892	4	6	-0.468	-0.1774	3	8	-1.006	-0.8039	6	4	-0.262	-0.0959	23	25	0	0.1566
3	17	4	-0.776	-0.4322	40	56	0.085	0.3287	6	17	-0.607	-0.4105	40	40	0.044	0.1660	42	21	0.297	0.4273
4	38	36	-0.186	0	59	55	0.561	0.7870	31	47	-0.210	0	53	61	0.273	0.3691	47	42	0.551	0.6704
5	53	57	0.146	0.2630	38	36	1.012	1.2395	62	59	0.495	0.4950	32	38	0.456	0.5350	31	43	0.789	0.9086
6	43	55	0.360	0.4425	22	18	1.476	1.7260	43	26	0.827	1.3085	32	17	0.608	0.6753	20	30	1.032	1.1628
7	13	21	0.513	0.5757	4	6	2.000	2.3113	26	20	2.288	-	5	11	0.738	0.7969	3	12	1.305	1.4658
8	3	4	0.632	0.6826	1	1	2.685	3.1771	-	-	-	-	1	3	0.852	0.9041	-	3	1.658	1.9089
9-10	1	2	≥0.683		2	-	≥3.177		-	-	-		2	3	≥0.904		1	2	≥1.909	

* r coefficient at the point of indifference between two adjacent lotteries/options.

Table D5 Mean and median values of r derived from DT16 – DT20

CRRA coefficient r	DT16		DT17		DT18		DT19		DT20	
	V1	**V2**	**V1**	**V2**	**V1**	**V2**	**V1**	**V2**	**V1**	**V2**
Mean	-0.097	0.185	0.690	0.610	0.643	0.362	0.311	0.319	0.500	0.661
Median	0.146	0.146	0.561	0.561	0.495	0.495	0.273	0.273	0.551	0.789

(b) Between-procedural comparison

- **The Allocation Procedure and Choice List Procedure**

In the Choice List Procedure, the median individual standard deviation of r values is equal to 0.420. For the allocation procedure, the corresponding statistic is 0.465. To abstract from such within-person variability and focus on an individual-level 'central tendency' measure, we take, for each individual, the median values of r generated by each procedure[28]. We then regress the medians from the allocation procedure (MedAll) on the medians from the choice list procedure (MedChoi). If the two procedures produce individuals' medians that are, on average, quite similar, we should expect the regression to yield an intercept close to zero and a slope that is not far from 1.

Because of the significant differences between V1 and V2 distributions for three of the five choice list tasks, we run separate regressions for each subsample. The results (showing standard errors in brackets below each coefficient) are:

$$\text{V1:}\quad \underset{}{\text{MedAll}} = \underset{(0.126)}{0.919} + \underset{(0.239)}{0.370}\,\text{MedChoi} \qquad R^2 = 0.014$$

$$\text{V2:}\quad \underset{}{\text{MedAll}} = \underset{(0.158)}{0.778} + \underset{(0.300)}{1.084}\,\text{MedChoi} \qquad R^2 = 0.069$$

In both cases the R^2 is low, but especially in the V1 regression, where the intercept is significantly greater than 0 and the slope coefficient is significantly less than 1. There is at best only a very weak relationship between the two measures here. In the V2 regression, the relationship is stronger – the slope coefficient is not significantly different from 1 – but the intercept is again significantly greater than 0, suggesting that the two elicitation procedures are liable to give substantially and systematically different CRRA measures[29]. Bearing in mind that we have tried to eliminate much of the within-procedure noise by using medians, the weakness and lack of agreement between the two sets of measures is discouraging.

[28] We take medians rather than means because some of the means were susceptible to influence from the occasional very high estimates of r generated by some people in some questions.

[29] The fact that these two regressions are themselves rather different from one another reflects the way in which the responses to the choice list tasks were affected by reversing the top-to-bottom order in those tasks.

- **The Ranking Procedure and Choice List Procedure**

To compare these two procedures we take individuals' median values of r produced by the choice list procedure and translate these into the decision each individual would make in the first ranking task, DT14 (which is closest to the E&G (2002) procedure). For example, as shown in Tables D2 and D3, someone for whom 0.6062 < r < 1 would rank (£20, 0.5; £6, 0.5) above all other options. By counting the numbers of individuals whose median r from the five choice list tasks lies in that range, we can infer the choice-list-based estimate of the frequency of those responses in DT14. And likewise for all of the other ranges of r.

Table D6 compares the actual distributions of responses to DT14 for each subsample with the inferred distributions. For both subsamples, there are very clear differences between the actual distributions and those based on the choice list procedure. Most strikingly, the choice list task *never* produces a median r high enough to entail ranking the sure £10 option first and only seldom produces a median r greater than 1, whereas the majority of individuals behave in the ranking task as if their values of r are greater than 1.

Table D6 Actual vs Inferred Top Ranked Options in DT14 Based on the CRRA Coefficients from the Choice List Procedure

	V1 Actual DT14	V1 Inferred DT14		V2 Actual DT14	V2 Inferred DT14
£10 for sure	70	0		77	0
£15, 0.5; £8, 0.5	28	5		22	9
£20, 0.5; £6, 0.5	23	25		26	26
£25, 0.5; £4, 0.5	13	62		5	54
£30, 0.5; £2, 0.5	2	39		4	35
£35, 0.5; £0, 0.5	36	40		45	55

- **Allocation Procedure and Ranking Procedure**

For these two procedures, we can repeat the comparison made in Table D6, but in this case we can infer the top ranked option from the median r derived from the allocation procedure. Since there were no significant differences between the subsamples in either of these procedures, we consider the pooled distributions in Table D7 below.

Table D7: Actual vs Inferred Top Ranked Options in DT14

	Actual DT14	Inferred from Allocation	
A: £10 for sure	147	10	
B: £15, 0.5; £8, 0.5	50	150	(75 + 75)
C: £20, 0.5; £6, 0.5	49	138	(63 + 75)
D: £25, 0.5; £4, 0.5	18	25	
E: £30, 0.5; £2, 0.5	6	17	
F: £35, 0.5; £0, 0.5	81	11	

Because 150 respondents had a median value of r = 1 in the allocation procedure, at which value someone would be indifferent between the second and third listed options, we have divided those individuals equally between each of those two options. However, it is clear that whatever way those 150 are divided, the actual distribution is significantly different from whatever might be inferred from the allocation procedure. In particular, the most extreme available options were actually selected by 228 (65.0%) of the sample whereas only 21 (6.0%) would have ranked them first on the basis of their median response to DT1-DT13.

The allocation procedure seems to pull responses towards r = 1, while the ranking procedure appears to encourage behaviour which results in more extreme (in both directions) values of r.

To sum up, we observe very weak consistency in the coefficients of CRRA within procedures and almost no consistency between procedures. Of course, these results may suggest that the CRRA specification is not a good way to represent participants' risk attitudes and that other functional forms might work better. Yet, our analysis presented in the main body of the paper makes us sceptical about the chances of success within the usual EUT set of options, because even the Spearman rank correlation coefficients, although statistically significant, are often quite low.